AFRL-RI-RS-TR-2013-176

FAUST: FLEXIBLE ACQUISITION AND UNDERSTANDING SYSTEM FOR TEXT

JULY 2013

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

AFRL-RI-RS-TR-2013-174 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/ S /	/ S /
JAMES M. NAGY Work Unit Manager	MICHAEL J. WESSING Deputy Chief, Information Intelligence Systems and Analysis Division Information Directorate

REPORT DOCUMENTATION PAGE	*Form Approved* **OMB No. 0704-0188**

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
JULY 2013	FINAL TECHNICAL REPORT	JUN 2009 – MAR 2013

4. TITLE AND SUBTITLE

FAUST: FLEXIBLE ACQUISITION AND UNDERSTANDING SYSTEM FOR TEXT

5a. CONTRACT NUMBER
FA8750-09-C-0181

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
62304E

6. AUTHOR(S)

L. Lynn Voss, David E. Wilkins, David Israel, Christopher Manning, Daniel Jurafsky, Daniel S. Weld, Pedro Domingos, Jude Shavlik, Christopher Ré, Andrew McCallum, David Smith, Michael Collins, Dan Roth, Eyal Amir, Cleo Condoravdi, Daniel G. Bobrow, Stanley Peters, Sriraam Natarajan, Yorik Wilks, Sergei Nirenburg

5d. PROJECT NUMBER
R674

5e. TASK NUMBER
00

5f. WORK UNIT NUMBER
02

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIEB
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TR-2013-176

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. PA# 88ABW-2013-3251
Date Cleared: 17 Jul 2013

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The vast majority of scientific and technical knowledge is expressed in natural-language (NL) texts. Our objective was to create an automated reading system that makes the knowledge in NL texts accessible to any of an open-ended range of formal reasoning systems. Our approach was based on large-scale statistical (probabilistic) joint inference over relational models. This vision involved a radical re-thinking of the architecture for Machine Reading systems.

15. SUBJECT TERMS
machine reading, joint inference, lifted inference, probabilistic relational models, text understanding, natural language process, NLP, NLU, NER, Named Entity Recognition, LDA, Latent Dirichlet Allocation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					JAMES N. NAGY
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPONE NUMBER (*Include area code)***
U	U	U	SAR	166	**N/A**

TABLE OF CONTENTS

SECTION **PAGE**

LIST OF FIGURES

FIGURE **PAGE**

LIST OF TABLES

TABLE **PAGE**

1. SUMMARY

1.1 BACKGROUND

The vast majority of scientific, technical, and expository knowledge is expressed in natural-language (NL) texts. In 2009, the Defense Advanced Research Projects Agency (DARPA) began the Machine Reading Program (MRP) to create an automated reading system that makes the knowledge in NL texts accessible to any of an open-ended range of formal reasoning systems.

Natural-language texts are written by and for humans, not for machine interpretation. They are both ambiguous (at the lexical and multiple structural levels) and inexplicit—they leave much crucial information unsaid. Their use depends on readers' abilities to disambiguate, make inferences, and supply missing information. Readers often do so by using prior information, generally gleaned from prior reading.

Previous natural language processing (NLP) research has addressed sources of ambiguity and many kinds of information gaps, but too often only as isolated problems. In reality, evidence that supports solutions to one ambiguity or information gap often comes from choices made with respect to others, and the evidence for and against each such choice typically interacts with evidence for and against the others. This observation motivated SRI International to assemble the only MRP reading team that was based on joint inference. Prior knowledge is also a source of evidence to be used during linguistic analysis, another benefit for a joint-inference approach.

1.2 HISTORICAL SKETCH

We begin with the historical context in which MRP was established. Near the beginning of systematic work in Text Understanding (in the 1970's), there were two major strands of research. There was a body of work, most strongly associated with Roger Schank, inspired by (cognitive) models of story-understanding in which hand-built, scenario-based templates were brought to bear (Schank and Colby 1973). This work, typically focused on hand-building "semantic (application-specific) grammars", was sufficiently limited in practical scope and overly brittle. Thus, even within that limited scope, it died a natural death. Analogues or descendants, however, can be found in two widely different arenas: First, the top-down, scenario- or task-driven approach can be seen in the quite vital and important work of the late 70's and 80's on task-oriented dialogues, which – with the spread of smart-phones – has taken on a new life in enabling transaction-oriented dialogues. The second arena was in the research, especially in the 90's, on manually built finite-state semantic grammars for closed-domain information extraction.

That other major strand of text understanding research, of more interest to Machine Reading, was what might be called The Classical (or: Good Old-Fashioned) Approach. This approach was built on top of large hand-written more-or-less application-independent "syntactic grammars", often quite directly inspired by linguistic theory. On the semantic side, meanings were expressed either in a logic-based language or in "AI Knowledge Representation" formalisms, e.g., KRL (Bobrow and Winograd 1977), or in one of the semantic-net formalisms. These latter were often "psychologically inspired" and, unlike the logic-based formalisms, lacked clear and systematic semantics. Here, the benefits of systematicity and high precision were overwhelmed by the cost in time of producing what were extremely limited applications.

Both of these approaches were largely swept aside by the statistical machine-learning revolution of the mid ‘90’s. The crucial prerequisites of this revolution were:

- Availability of large annotated data-sets and huge quantities of unlabeled text data.
- Moore’s Law; huge advances in memory capacity, processor speed, etc.
- Growth of the practice of community-wide open evaluations and of a metrics-focused research community.

The key concrete result of the revolution is that large parts of the NLP community moved toward building atheoretical, statistically-trained, ML-induced NLP modules (e.g., Part-of-Speech Taggers, Named Entity Extractors, Semantic Role Labelers, Parsers). Until fairly recently, a corollary was that sentence or clause-level semantics were almost completely ignored.

With this little bit of historical context, we can say the SRI team’s vision was to unify what was worth saving in the Classical Approach to text understanding with the best-of-breed of the statistically-trained machine-learning revolution in natural language processing.

1.3 ARCHITECTURAL SKETCH

There are many different paths towards unification. Figure 1 gives us one way of thinking about these paths by presenting an admittedly simplified picture of architectural options. Candidate architectures include Monolithic, Annotation Pipeline, and Loosely Coupled Components. The horizontal axis is the degree to which one opts for independent modules in putting together an end-to-end MR system. IBM’s Watson (Ferrucci et al. 2010), or more generally its Deep QA system, is an example of a choice to build a system based on a large number of independent modules that are black boxes, one to another. On the vertical axis is the degree to which the system makes systematic use of evidence from a wide variety of sources. The yellow diamond at the bottom right represents Watson in some respects, but this diamond could migrate up the vertical axis. These two “axes” are not really independent, of course; but they are less co-dependent than one might think. Thus, in a case like Watson’s, there could be modules that, acting in concert, ensure all the available evidence was used before a final action (e.g., generating an answer to a query, or in Watson’s case, generating a query to an answer). Indeed, something like this happens in Watson’s case, as we understand it, with the responsibility being shared between question-decomposition at the front-end and final decision-making (with a confidence threshold) at the back-end.

A useful contrast with a system like Watson is provided by what we will call Hobbs’ One Big Engine picture, a version of which is described in “Interpretation as Abduction” (Hobbs, et al. 1993). In that picture, there are no modules; hence no NLP modules. There is not a Part-of-Speech Tagger or a Semantic Role Labeler or a Named Entity Extractor or a Parser. Rather there are formulations (axiomatizations) of the principles that govern the behavior of such modules, all expressed in the same language (let us say, full first-order logic, perhaps with extensions) that one uses to express domain knowledge (and knowledge of the domain-independent structure of discourse, etc.). In the 1993 paper, the matrix predications in these formulations all get abductive weights and there is a single uniform abductive procedure – which in special cases, reduces to standard deduction – that results in a “theorem” formulating the best overall interpretation of a given passage of a text.

Given these two dimensions of architectural options, the SRI team can be seen as having followed a modular approach, at least to the extent of having chosen, unlike the One Big Engine, to build upon existing NLP-modules. On the other hand, we have also partially adopted the One Big Engine picture, at least to the extent that all these modules communicate via a common language, the language of Markov Logic, which is also the language in which domain-knowledge (typically probabilistically characterized) is expressed. Moreover, there is a single uniform inference procedure for fusing all this probabilistic information, thereby enabling joint inference across all the modules in the system. V1 represents our initial system, and v5 represented our goal at the end of Phase 5 of MRP.

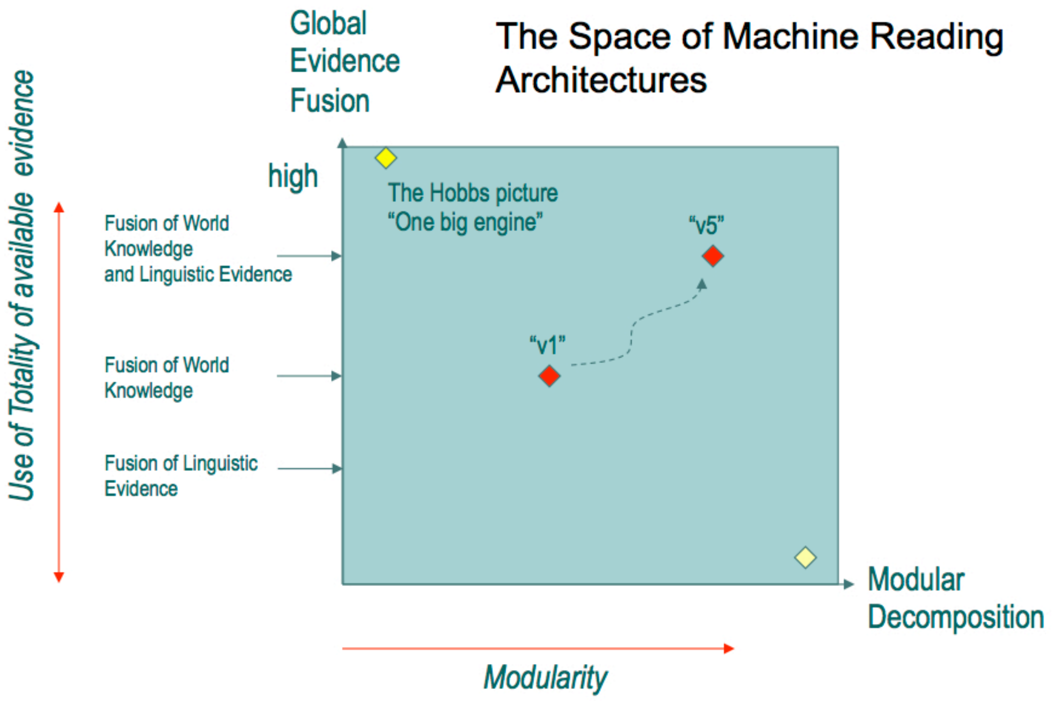

Global
Evidence
Fusion
The Space of Machine Reading
Architectures
Use of Totality of available evidence
high
Fusion of World Knowledge and Linguistic Evidence
Fusion of World Knowledge
Fusion of Linguistic Evidence
The Hobbs picture
"One big engine"
"v5"
"v1"
Modular
Decomposition
Modularity

Figure 1: Space of possible architectures for doing machine reading

SRI International's vision involves a radical rethinking of the architecture for MR systems. Our use of statistical (probabilistic) joint inference over relational (first-order) models as the core technology represents a very different approach from other MR systems, including the two other reading systems funded under MRP.

1.4 SUMMARY OF RESULTS

MRP lasted for only three of the planned five phases, so we did not get to fully test our overarching hypothesis that a machine reading system based on joint inference (JI) could be made tractable. At project start, the current JI methods were not robust enough to perform on the scale required by our vision. However, we did make considerable progress, and results so far are, at the very least, promising.

To realize our distinctive research vision, we assembled a team that included leading researchers in NLP, probabilistic reasoning, and machine learning (ML). Our progress toward our vision is made evident by our team's extensive contributions to scientific knowledge. FAUST researchers have won five best paper awards and have published 155 papers (so far), most in top conferences, for their papers on MRP-sponsored work (see Appendix A). In addition, the FAUST team developed extensive MR-related software that is freely available (see Appendix B).

A key for achieving our goals was fully supporting the software integration tasks and MRP evaluations. We had both an excellent and experienced Software Engineering (SE) team and sufficient funding for the integration and evaluation tasks, which are often underestimated. Helping the Government define the evaluations, preparing for them, and participating in them consumed a significant amount of our effort. The Government Evaluation Team (ET) will report on the evaluations and their results, so this report will concentrate on our research results.

Our team made major advances and explored new directions in NL understanding, at levels ranging from providing general infrastructure components useful to many groups to cutting-edge research into new models of language. A key result was the development of Stanford's CoreNLP, a simple-but-flexible pipeline framework that ties together all of Stanford's core NLP components, from sentence splitting and tokenization through parts-of-speech, named entities, to parsing and co-reference, and makes them available under a simple uniform API. CoreNLP was made publicly available open source. To add temporal information to CoreNLP, Stanford created SUTime, a Java library that recognizes and normalizes temporal expressions using deterministic patterns [101].

UIUC made another such framework available to the research community — their Curator is a distributed system for running and aligning multiple-state NLP preprocessing tools, as well as state-of-the-art tools for multiple NLP tasks, including semantic role labeling, named entity recognition, and co-reference resolution [103].

Stanford developed a new deterministic sieve architecture for entity co-reference. This system was the best performing system at the CoNLL 2011 Shared Task [68] on entity co-reference. The FAUST team developed improved relation-extraction systems. This was explored using both fully supervised methods over linguistic analyses, as in the Phase 2 evaluation, and more extensively by considering the task of distantly supervised learning. Stanford worked on this in the context of the NIST TAC KBP task, and developed a new, principled model that handles the uncertainties of distantly supervised learning (the MIML-RE model).

Stanford explored the usage of joint learning methods within NLP. They extended work on co-reference and event extraction, introducing a new model of cross-document joint entity and event co-reference. They initiated a major exploration of deep learning (multi-layer neural network) methods for use of the data-dependent recursive hierarchical structures of natural language. A paper on this work won an ICML 2011 best paper award [70].

The University of Wisconsin (Madison) developed modules for very-large-scale JI.—Tuffy (a Markov Logic network RDBMS-based inference engine [48,87], which has been downloaded more than 5000 times) and Felix (an operator-based relational optimizer for statistical inference) [85]. They developed new approaches for very-large-scale inference, including optimization approaches such as dual decomposition and partitioning-based inference algorithms. Wake Forest and SRI developed the Anytime Lifted Belief Propagation (ALBP) algorithm. Wisconsin demonstrated that their approach of using probabilistic logic to extract information from text scaled to a corpus of more than one billion documents. Wisconsin and Wake Forest developed novel approaches and algorithms to address several open problems in Statistical Relational Learning (SRL). These approaches were quite effective when applied to MR and other text-based datasets [93].

The University of Washington (UW) pioneered and extended a diverse set of approaches for distant supervision of relational extractors. Their methods used background knowledgebases ranging from Wikipedia, Freebase, and the Nell KB, and matched to a variety of textual corpora including Wikipedia and newswire text. Their LUCHS system generated extractors for more than 5000 distinct relations [21], which is several orders of magnitude more than previous systems. Their VELVET system introduced the notion of ontological smoothing, a method for quickly training a relational extractor with only a handful of positive examples [137]. UW implemented and evaluated the Ontological Belief Propagation algorithm [78].

UW also worked on unsupervised semantic parsing and enabling efficient large-scale JI. UW created new architectures and algorithms for efficient, large-scale probabilistic JI; (2) developed algorithms that unify probabilistic and logical inference; and (3) developed methods for scalable semantic parsing from text [1,43,72,73,74,78,84].

The University of Massachusetts Amherst developed a joint model for event extraction that combined entity-type prediction and detection of event arguments [76]. UMass built the first cross-document joint NER and relation-extraction model, trained only with weak supervision [35]. UMass developed SampleRank [12], a highly scalable algorithm for learning in large-scale graphical models and pioneered a new paradigm for distant supervision by introducing latent variables that indicate whether a relation is expressed by a mention. This paradigm has improved the accuracy of relation extraction. UMass developed several algorithms to make JI scalable and they were orders of magnitude faster than previous methods. The speed was achieved by lazily instantiating both factors and variable values only when they were needed. UMass developed a new generation of cross-document co-reference algorithms that rely on hierarchies of co-reference clusters for both increased robustness and efficient parallel inference.

SRI and UMass developed a general framework for lifting variational approximation algorithms such as linear programming relaxation of maximum *a posteriori* (MAP) inference [150], a widely used approximation in NLP problems.

UIUC pioneered an ILP-based framework to support incorporating declarative knowledge as a way to guide learning and support global inference [63]. They developed new algorithms for learning with indirect supervision, and for learning and inference with latent representations. The UIUC framework was used in developing multiple NLP capabilities, including (1) relation and event extraction; (2) co-reference; (3) textual inference; and (4) temporal and causal reasoning [130,133]. They developed the Wikifier, an approach for disambiguating concepts and entities appearing in text and grounding them in an encyclopedic resource. UIUC (Prof. Amir) worked on probabilistic modal (PM) operators for natural language understanding. They investigated using PM models to represent what authors assume about readers' knowledge, and created both a theoretical framework for inferring Bayesian Network PM models from text and an implementation.

UIUC and UMass jointly developed an abstraction of a joint-inference module that will enable integrating individually learned modules and running multiple forms of joint inference for it.

PARC and CSLI's work focused on inferences that can be drawn from texts based on inferential properties of linguistic expressions. They demonstrated that the task can be aided by different kinds of resources, including lexical class markings, ontological classifications, and domain models that link different classes of items together. They developed an analysis of the range of veridicality signatures and the environments in which lexical items have these signatures, and verified this analysis with human subjects. They developed an algorithm of projection that took into account the effect of contextual factors.

MIT/Columbia developed novel methods based on dual decomposition and Lagrangian relaxation for inference in NLP. The method was shown to be effective in a number of NLP problems. The work resulted in several publications, including a best paper award at EMNLP 2010 [36]. They developed novel spectral-learning methods for latent-variable models, and completed experiments showing that the methods perform at the same level of accuracy as Expectation Maximization (EM), which is widely applied in NLP, but are an order of magnitude more efficient in training time.

IHMC executed an exploratory project in Phase 3R to locate proto-beliefs of individual Ummah message board posters on a large scale. Facts were extracted from the Ummah message board postings using unsupervised methods for information extraction. These facts were then linked to individual posters as beliefs or assertions in a belief management engine. The primary outcome of the completed work is a positive demonstration of the extraction of these beliefs.

Ellipsis is a linguistic process that renders certain aspects of text meaning invisible at the surface structure. Ellipsis is considered one of the more difficult aspects of text processing and, accordingly, has not been widely pursued in NLP applications. Onyx worked in Phase 3R toward a system that can resolve one class of elliptical phenomena: elided scopes of modality. Onyx developed a full microtheory of modal-scope ellipsis treatment [151] and a method of detecting and resolving a subset of cases of modal scope ellipsis that can be applied to big data.

SRI's research team developed *Lifted probabilistic inference,* which manipulates the representation in first-order form, keeping it compact and performing operations on a single conditional probability function. By contrast, regular inference would perform the same operations repeatedly, for each instance of that function [138,139]. SRI developed an engine for *anytime lifted probabilistic inference,* an incremental inference method that updates a query's answer gradually as it examines increasingly relevant portions of the model. If the query depends only on a small fraction of the model, as most do, then the algorithm will not examine the entire model to find the answer [141]. SRI developed a new lifted inference method, LIDE (Lifted Inference with Distinct Evidence) that allows polynomial-time exact lifted inference [138].

2. INTRODUCTION

2.1 APPROACH

SRI International's vision—in which large-scale statistical (probabilistic) joint inference over relational (first-order) models is the key technology—represents a very different approach from other Machine Reading (MR) systems. Our vision involves a radical rethinking of the architecture for MR systems. We reject the standard pipeline approach, in which lower-level decisions (e.g., as to part of speech) are finalized before being passed to higher-level processing modules (such as parsers). Moreover, pipelines tend to have no place—hence little use—for non-linguistic information. In contrast, lower-level analyses in our approach are treated as partial and probabilistically weighted, and our approach enables probabilistic inferences over these partial and uncertain analyses by using a wide variety of non-linguistic information sources, which are also probabilistically weighted.

SRI's reading system is FAUST, the Flexible Acquisition and Understanding System for Text. FAUST implements innovative solutions to the key challenges that arise when bridging from knowledge encoded in natural language to knowledge for use by computational reasoning systems. The FAUST architecture is based on *statistical joint inference over probabilistic relational (first-order) models.* FAUST supports the simultaneous consideration of multiple random variables and the relationships among them. The latter are explicitly expressed in a rich, probabilistic, first-order language (Markov Logic). Theories in this language are converted into probabilistic, relational (factor) graphs and then probabilistic inference algorithms are run over them. This architecture and key technology enable FAUST to:

- Leverage a wide range of mutually constraining information (both linguistic and extra-linguistic evidence) for decisions at any level of analysis, all expressed in a single common language
- Integrate information across multiple sentences and texts.

Our team explored a range of joint-inference engines that integrate information from an ensemble of state-of-the-art NLP modules provided by our team's NLP-focused researchers (see Section 2.1, which describes the team members' roles). These modules provided the linguistic information (e.g., as to parts of speech, semantic role, syntactic structure); that is, they generate probabilistic, linguistic evidence by analyzing the text. This information is then combined with domain-specific knowledge, including rules ("bridge rules") that relate objects, relations, and events in a domain with the linguistic features of texts containing information about that domain. Such cross-level and cross-source integration requires *aligning* representations from both non-linguistic sources and from multiple levels of textual analysis.

FAUST attempts to continually improve these alignments by using machine-learning techniques. Because NL texts introduce new concepts and support new generalizations but seldom contain explicit definitions of these concepts or generalizations, new reasoning methods are required. As such, we enabled the learning of both new concepts and new generalizations from NL text by developing a set of concept- and rule-induction mechanisms. Crucially, we used these to refine what was previously learned. Further, by making available the joint-inference results to each module as "noisy" training examples, we enabled continuous improvement in FAUST's reading ability; more specifically, we enabled using the knowledge acquired from texts to improve the very process of acquiring knowledge from further reading.

This vision of tightly integrated NLP, probabilistic reasoning, and a learning system—both to enable improved reading and to acquire domain knowledge by inference from what is acquired by reading—was fully supported by the expertise and experience of our team. SRI's team included leading researchers in NLP, probabilistic representation and reasoning, and machine learning as well as software engineers with extensive integration experience on similar projects.

Our approach had risk, as we counted on making scientific breakthroughs in applying probabilistic reasoning on such a broad scale. The risk was mitigated as we were building on significant recent advances in probabilistic representation and joint inference. Further, under MRP, we have defined a Joint-Inference API that enables widely varying FAUST modules to share probabilistic hypotheses and reasoning. The resulting benefits justified the risk:

- The FAUST system considered the widest range of both linguistic and extra-linguistic evidence (to achieve better results), using the same mechanisms that are applied for the integration of linguistic information across multiple levels.
- Our team made extensive contributions to scientific knowledge. Our world-leading researchers won multiple best paper awards and published 155 papers, most in top conferences, for their papers on MR-sponsored and MR-related work. These papers are listed in Appendix A, and the citations by number in this document refer to the papers in Appendix A.
- The FAUST team developed extensive MR-related software that is free and openly available. These software modules can be found in Appendix B, with instructions for downloading, when applicable. Most software modules mentioned in this report can be found in Appendix B.

To summarize, the FAUST team attempted an admittedly high-risk venture: we believe that *intelligent* machine reading requires the radical rethinking sketched above, which imposes new architectural requirements that Team FAUST explored. Our animating vision was that truly significant progress in machine reading requires that we must do more than advance the state-of-the-art in the component NLP technologies—we must advance the science of combining the outputs of the various modules that embody those technologies with various kinds of non-linguistic information. Further, we believe that by far the most promising way to do that is large-scale joint probabilistic inference over relational models.

2.2 PROJECT OBJECTIVES

The goal of the Machine Reading Program (MRP) was to automatically make the knowledge contained in natural-language (NL) texts accessible by formal reasoning systems. Real examples of such formal systems include (Relational) Database Systems; Bayes Net reasoners; Datalog/Logic Programming systems; OWL and other Description Logic systems; systems that reason with first-order languages (such as theorem provers); Probabilistic Database systems; and systems that reason with Probabilistic Relational Languages.

This goal presented four key technical challenges:

(1) Capturing and representing the information needed to determine the meaning of texts (a) from multiple linguistic levels (corpus, document, paragraph, sentence, clause, phrase, and word levels), and (b) from knowledge beyond the text, including knowledge acquired from reading other texts.

(2) Dealing with the uncertainty of heterogeneous interpretative hypotheses that arise from reading, and interrelating these hypotheses so that their interdependences can appropriately constrain the evolving interpretation choices.

(3) Efficiently reasoning with this uncertainty and heterogeneity on a large scale.

(4) Improving system reading performance by learning from reading and learning for reading.

As with most multi-year research programs, we had two overarching goals: (1) to increase the rate of progress along pre-existing trajectories of the relevant state-of-the-art, and (2) to move the curve out, changing these trajectories for some of the technologies. More specifically, we set as goals the following innovative solutions to those overarching aims and derivative challenges, as stated prior to the project's start:

- **Joint Inference for NLP.** We propose a radical re-thinking of the architecture for NLP tasks. Instead of the standard pipeline approach, FAUST implements a model centered on extended joint inference over probabilistic relational models, which enables generating and aligning representations at multiple levels of analysis, and thus leveraging all mutually constraining information expressed in them.
- **Management of Uncertainty.** We will build on our team's groundbreaking work and extensive experience in practical, large-scale probabilistic joint inference. We will develop a variety of localization, factoring, and approximate inference techniques so that FAUST can efficiently harness sub-ensembles of tightly linked information sources, independently of the entire ensemble of potential sources.
- **Integration of Information across Sentences and Texts.** Most NLP research has focused on the sentence as the unit of meaning; but information in natural texts is not localized in sentences. Rather, it is distributed across larger discourse units; indeed, multiple distinct texts will discuss the same items and concepts. FAUST will interpret discourse relations, gather distributed information, and use sophisticated inference over partial representations to integrate this information into one coherent model.

- **Use of Extra-Linguistic Knowledge.** Authors assume human readers will use prior information, often gained from prior reading. Such knowledge is also a source of constraints on the choice of linguistic analysis of the current text. While current probabilistic NLP can incorporate such constraints in the form of co-occurrence statistics, our new knowledge-aware NLP architecture will enable FAUST to leverage the full range of evidence, both linguistic and extra-linguistic, to support intelligent reading.
- **Learning New Concepts and Rules by Reading.** Human readers use what they learn from texts as a basis for further inference-based learning. To emulate this crucial ability, FAUST will support learning both new concepts and new generalizations from natural text, and will refine previously learned knowledge, with an ensemble of concept- and rule-induction mechanisms.
- **Continuous Improvement in Reading.** Humans improve their ability to read by checking their understanding of texts in a variety of ways, including seeking confirmation (or experiencing disconfirmation) by further reading. The joint inference approach in FAUST will enable the system to use previously learned knowledge to improve reading performance.
- **Adaptation for New Domains.** One of the major constraints on building highly capable Machine Reading systems is and will continue to be the scarcity of labeled data. We will use a wide variety of machine-learning techniques, coordinated through joint inference, to learn from the large quantities of easily available naturally occurring data, and thus to adapt to new domains and new tasks.
- **Support for an Open-Source Research Community.** Finally, an added and critical benefit of our joint inference approach is that it will support the growth of a plug-and-play open-source research community by providing a uniform approach to introducing new reasoning and learning modules.

3. METHODS, ASSUMPTIONS, AND PROCEDURES

A crucial element of our approach is that reading, especially to extract information for reasoning systems, requires an integrated, comprehensive set of NLP modules and significant reasoning capabilities. The system must reason to (1) integrate the inputs from its NLP modules to synthesize the most coherent and probable total interpretations of the texts it reads, and (2) make further inferences from those interpretations to continuously enhance its ability to extract useful information when reading.

3.1 FAUST TEAMING STRATEGY

One key method for achieving our goals was selecting the best research team to realize our distinctive research vision: Intelligent reading of natural-language texts requires reasoning, especially reasoning that takes account of both linguistic and nonlinguistic sources of information. The system must reason both to integrate a variety of inputs (some nonlinguistic) to synthesize the most coherent and probable total interpretations of the texts, and to make further inferences from those interpretations to continuously enhance the extraction of useful information. To achieve this vision, we assembled a team that includes leading researchers in NLP, probabilistic reasoning, and machine learning (ML). Figure 2 depicts our team members and their organization for the first two phases of the project. In Phase 3, there were some additions and some researchers moved to new institutions and continued on the team

NLP and ML: *Stanford University* (Chris Manning and Dan Jurafsky); *University of Massachusetts, Amherst* (Andrew McCallum, Sebastian Riedel and David Smith); *University of Illinois, Urbana Champaign* (Dan Roth); *Columbia University (formerly MIT)* (Michael Collins); and PARC/*Stanford*'s *Center for the Study of Language and Information (CSLI)* (Daniel Bobrow, Cleo Condoravdi, Annie Zaenen, and Lauri Kartunen) provided FAUST with state-of-the-art NLP modules and wide-ranging expertise. Stanford's CoreNLP system was the backbone of our efforts to determine the semantically relevant linguistic structures and features of sentences. UMass worked closely with Stanford, using Stanford's CoreNLP system as the basis for experimenting with a variety of joint-inference schemes. UIUC contributed a second end-to-end NLP system, with a different architecture and a special focus on the use of nonlinguistic information as a source of constraints on linguistic processing. CSLI focused on lexical semantics and on the linguistic expression of time and temporality. Columbia explored a variety of regimens for probabilistic joint inference among NLP modules. The FAUST NLP team also included the *University of Washington* (Dan Weld), which exploited a range of information sources to support quick adaptation to new relations and new domains.

Joint Inference (JI): Our vision required significant advances in large-scale probabilistic joint inference over relational models. The *University of Wisconsin* (Jude Shavlik and Chris Ré) and *SRI* (Hung Bui and Rodrigo de Salvo Braz) each contributed probabilistic inference systems. Both built on the work on Markov Logic Networks (MLNs) of another FAUST teammate, the *University of Washington* (Pedro Domingos), which explored probabilistic theorem proving to provide a unified approach to JI by combining lifted inference and sampling. Wisconsin extended probabilistic database technology to handle the very large numbers of ground facts that result from the application of MLN theories of textually encoded information to real texts. SRI extended its Probabilistic Consistency Engine to handle much larger-scale networks, while also developing a new scheme for large-scale lifted probabilistic inference (that is, inference at the level of quantified, not fully instantiated information). These efforts exploited the fact that the FAUST team provided special modules and algorithms to perform the common NLP subtasks of Machine Reading (e.g., named-entity extraction, relation extraction, and co-reference resolution).

NLP and JI: Our NLP teammates, especially those from UMass, UIUC, and Columbia contributed expertise on JI, and worked with Wisconsin, UW, and SRI to design and implement our cross-module JI APIs. We aimed to develop the most suitable and efficient mechanism for interfacing between FAUST's NLP modules and its multiple probabilistic JI engines.

During Phase 3, there new research teams were added to our team at DARPA’s request. These teams were Institute for *Human and Machine Cognition* (IHMC, Yorick Wilks), *Onyx Consulting* (Sergei Nirenburg), and Eyal Amir’s group at *UIUC*.

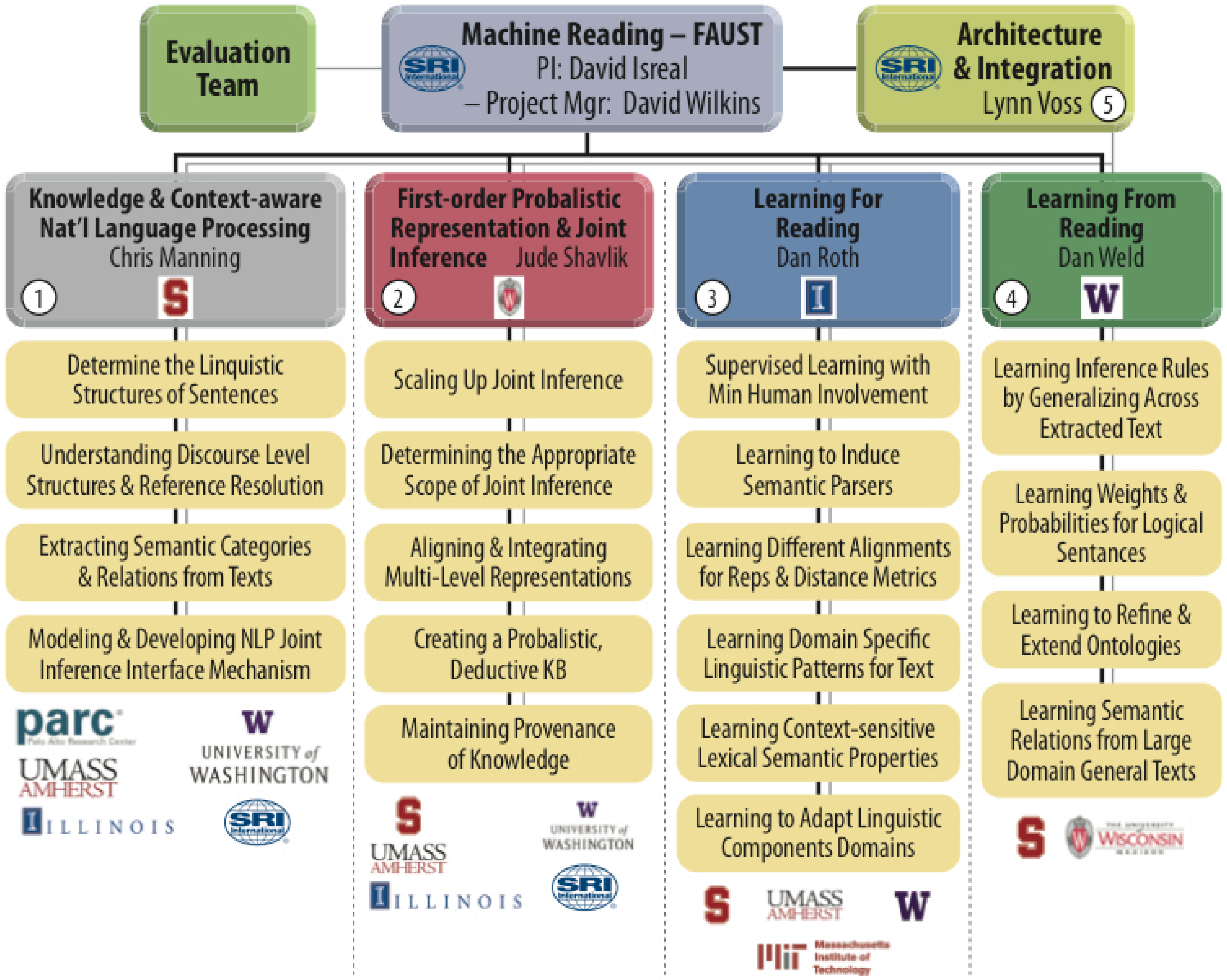

Evaluation Team
Machine Reading – FAUST
PI: David Isreal
– Project Mgr: David Wilkins
Architecture & Integration
Lynn Voss 5
Knowledge & Context-aware Nat’l Language Processing
Chris Manning
1
First-order Probalistic Representation & Joint Inference
Jude Shavlik
2
Learning For Reading
Dan Roth
3
Learning From Reading
Dan Weld
4
Determine the Linguistic Structures of Sentences
Understanding Discourse Level Structures & Reference Resolution
Extracting Semantic Categories & Relations from Texts
Modeling & Developing NLP Joint Inference Interface Mechanism
Scaling Up Joint Inference
Determining the Appropriate Scope of Joint Inference
Aligning & Integrating Multi-Level Representations
Creating a Probalistic, Deductive KB
Maintaining Provenance of Knowledge
Supervised Learning with Min Human Involvement
Learning to Induce Semantic Parsers
Learning Different Alignments for Reps & Distance Metrics
Learning Domain Specific Linguistic Patterns for Text
Learning Context-sensitive Lexical Semantic Properties
Learning to Adapt Linguistic Components Domains
Learning Inference Rules by Generalizing Across Extracted Text
Learning Weights & Probabilities for Logical Sentances
Learning to Refine & Extend Ontologies
Learning Semantic Relations from Large Domain General Texts
parc
UMASS AMHERST
UNIVERSITY of WASHINGTON
ILLINOIS
Massachusetts Institute of Technology
WISCONSIN

Figure 2: The FAUST team for Phases 1 and 2, and their task responsibilities.

3.2 SOFTWARE ENGINEERING AND INTEGRATION

A second key method for achieving our goals was fully supporting the software integration tasks and Machine Reading Program evaluations. This support included both an excellent and experienced SE team and sufficient funding for the integration and evaluation tasks, which in our experience are often underestimated.

SRI's Software Engineering team used its experience on several large DARPA programs to successfully design and integrate FAUST. On a wide range of DARPA and IARPA projects, our SE staff works with all parties to bridge the gap between novel, often university-based research and the government’s goal for mature, usable technology. In particular, SRI has significant experience integrating probabilistic and adaptive learning components (key technologies in our MR approach) from different institutions into a single, consistent architecture. SRI is also experienced in testing and validating early prototype components, and in testing, documenting, and delivering working systems to both DARPA and other clients. SRI is experienced with the complexities of formal evaluations conducted by independent evaluation teams.

Our procedure in the MR program was for SRI to exploit and continue to refine its repertoire of large-system integration tools, experiences, and practices, while avoiding the mistakes of past programs. This repertoire has been assimilated from many significant software integration projects involving university subcontractors, including the Bootstrapped Learning, GALE, and PAL/CALO DARPA programs that our SE team was recently involved with before Machine Reading.

The SRI Software Engineering team planned, managed, and executed all software integration tasks and MR program evaluations for the FAUST system. Helping the Government define the evaluations, preparing for them, and participating in them consumed a significant amount of our effort. Our lessons learned during these and similar projects increased the likelihood of our success on MRP. Our integration plans were built to identify and eliminate risk.

SRI avoided the risky “fix it later” approach by setting up our state-of-the-art software-engineering-infrastructure tools and processes at the start of the program, including automatic builds; regression tests; issue identification, tracking and resolution; and multiple methods for enhancing quick-reaction team collaboration and reducing technical risk. SRI began working with our team members early assigning individual liaisons to closely collaborate with each subcontractor to define and use APIs and functional requirements, and to ensure the correct use of SRI’s processes for software production.

Finally, a benefit of SRI’s SE methodology is that the promising technologies developed by SRI and our subcontractors will be more valuable to future DARPA programs and to the research community in general; their modules will be more easily reused in future NL understanding programs because of the better software engineering, documentation, and testing.

4. RESULTS AND DISCUSSIONS

4.1 PROJECT PHASING

The original plan for Machine Reading as described in the BAA was to have five phases, where each phase focused on a different set of goals that contributed to the overall vision. Phase 1 focused on an evaluation of readability assessment. As executed, Phase 1 was shortened by DARPA. Both the length and the type of evaluation made Phase 1 significantly different from the later phases. In 2012, DARPA decided to end the MR program after Phase 3, and, under a new Program Manager, Dr. Bonnie Dorr, decided to refocus the Phase 3 efforts in mid-phase toward research and away from hardening and evaluating an end-to-end system for Machine Reading.

We will therefore describe Phase 3 in two parts, which we will call Phase 3 and Phase 3-Research (3R). Phase 3R is so named because DARPA gave explicit guidance to not use resources to produce an end-to-end system in this phase, but instead to concentrate on research goals. Each institution on our team submitted a white paper to DARPA describing a detailed plan for their remaining Phase 3 research and associated funding levels for DARPA’s approval. In addition, DARPA invited two new teammates, as well as an additional research team from one of our existing subcontractors, to submit white papers. All three were added to our team.

Thus, Phases 1, 2, and 3 each had unique character, and we therefore organize this report by Phase. We define the temporal extent of these phases as follows:

- **Phase 1:** From 6/4/2009 to 3/31/2010. SRI got under contract on 6/4/2009; however, many subcontractors were not under contract until two months later, as DARPA requested that our effort be minimal until the Kickoff Meeting, which was 8/10/09. This date marks the true start for most of our team. The evaluation of readability assessment, which was the Phase 1 evaluation, occurred during the first quarter of 2010.
- **Phase 2:** The Kickoff meeting for Phase 2 started on 4/10/2010. Phase 2 covered the last three quarters of 2010 and the first quarter of 2011. Our team created an end-to-end Machine Reading system for the Phase 2 evaluation, which started in January of 2011.
- **Phase 3:** The Kickoff meeting for Phase 3 started on 4/5/2011. Phase 3 covered the last three quarters of 2011. We retested part of the Phase 2 evaluation in May of 2011 and did part of the planned Phase 3 evaluation in November 2011.
- **Phase 3R:** We consider Phase 3R to cover all of 2012, and for three institutions, the first quarter of 2013. Some subcontractors finished their work before the final quarter of 2012; most finished it by the end of 2012; and SRI, Wake Forest University, and Stanford University’s CSLI finished during the first quarter of 2013.

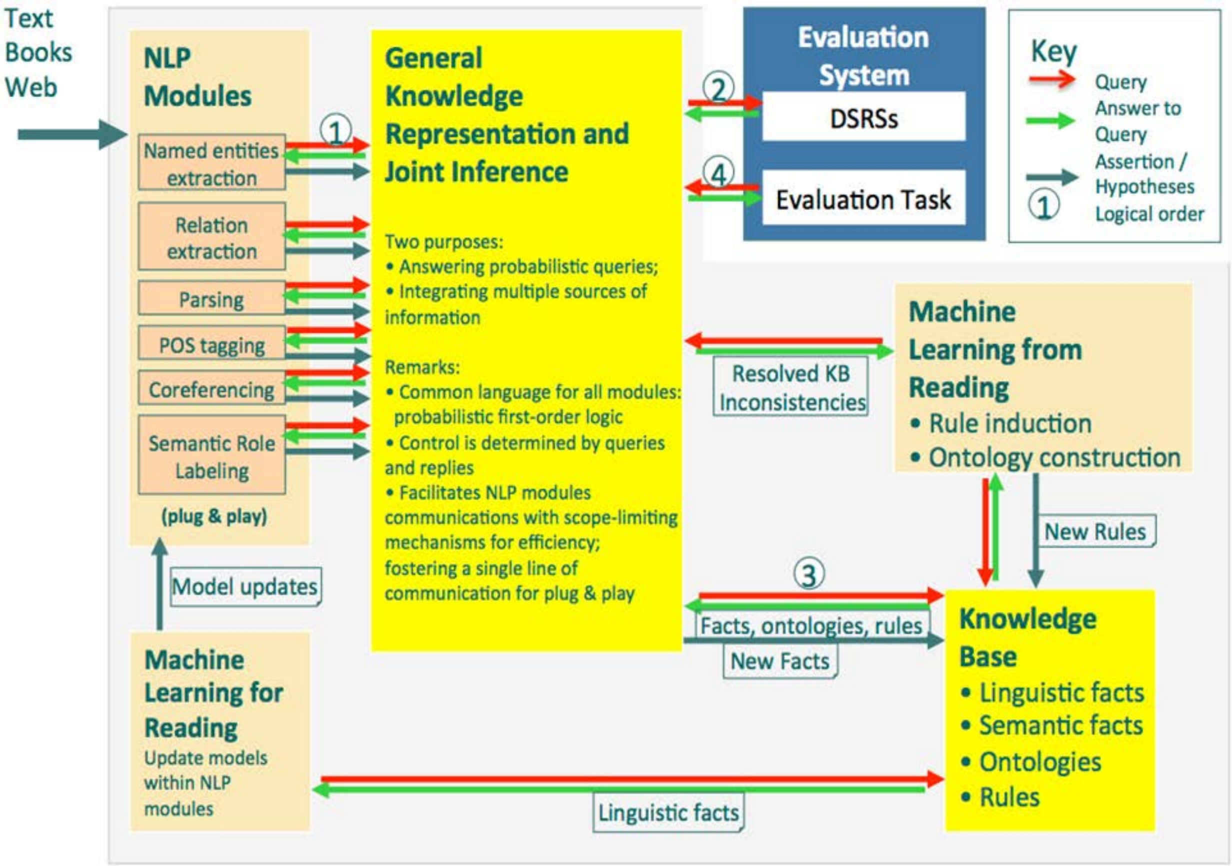
In the remainder of this section, we summarize the work performed and the major results for each of these phases of the FAUST project.

4.2 OVERVIEW OF RESULTS, ALL PHASES

The FAUST team made extensive contributions to scientific knowledge. Our world leading researchers won multiple best paper awards and published 155 papers (so far), most in top conferences, for their papers on MR-sponsored and MR-related work. These papers are listed in Appendix A, and the citations by number in this document refer to the papers in Appendix A.

The FAUST team developed extensive MR-related software that is free and available. These software modules can be found in Appendix B, with instructions for downloading.

As described in Section 1, FAUST chose a Loosely Coupled Components approach, applying "plug-and-play" NLP components. All components would communicate with a substrate of statistical relational learning, enabling joint inference across the system. Figure 3 shows the FAUST system, and the large yellow box in the top center provides a pathway from individual NLP components to perform complex, joint inference across components. Domain-specific reasoning systems (DSRSs), provided by the Government Evaluation Team, were used in the evaluations.



Text
Books
Web
NLP Modules
Named entities extraction
Relation extraction
Parsing
POS tagging
Coreferencing
Semantic Role Labeling
(plug & play)
Model updates
Machine Learning for Reading
Update models within NLP modules
General Knowledge Representation and Joint Inference
Two purposes:
• Answering probabilistic queries;
• Integrating multiple sources of information
Remarks:
• Common language for all modules: probabilistic first-order logic
• Control is determined by queries and replies
• Facilitates NLP modules communications with scope-limiting mechanisms for efficiency; fostering a single line of communication for plug & play
1
2
3
4
Evaluation System
DSRSs
Evaluation Task
Key
Query
Answer to Query
Assertion / Hypotheses
1
Logical order
Resolved KB Inconsistencies
Machine Learning from Reading
• Rule induction
• Ontology construction
New Rules
Facts, ontologies, rules
New Facts
Knowledge Base
• Linguistic facts
• Semantic facts
• Ontologies
• Rules
Linguistic facts

Figure 3: FAUST System architecture

When necessary, various components were more closely joined into subassemblies for efficiencies. We note that the two other candidate architectures can be cast into this architecture. This approach offers the following advantages:

- Modularity (important for engineering; see annotation pipeline)
- Chance to overcome cascading errors
- Relatively efficient (if the components partition model into tractable substructures)
- Theoretically sound

However, it has some disadvantages. It is slower than an annotation pipeline, the components need to “predict their inputs”, and it poses a software-engineering challenge.

The Stanford Core NLP Pipeline shown in Figure 4 was used by most of the members of the FAUST team as the NLP Modules (see Figure 3) that initially ingest text. UIUC contributed a second end-to-end NLP system, with a special focus on the use of nonlinguistic information as a source of constraints on linguistic processing. Instead of producing one annotated parse, it produced a set of possibilities with probabilities, so that latter information could be used to update the probabilities and pick the correct alternative.

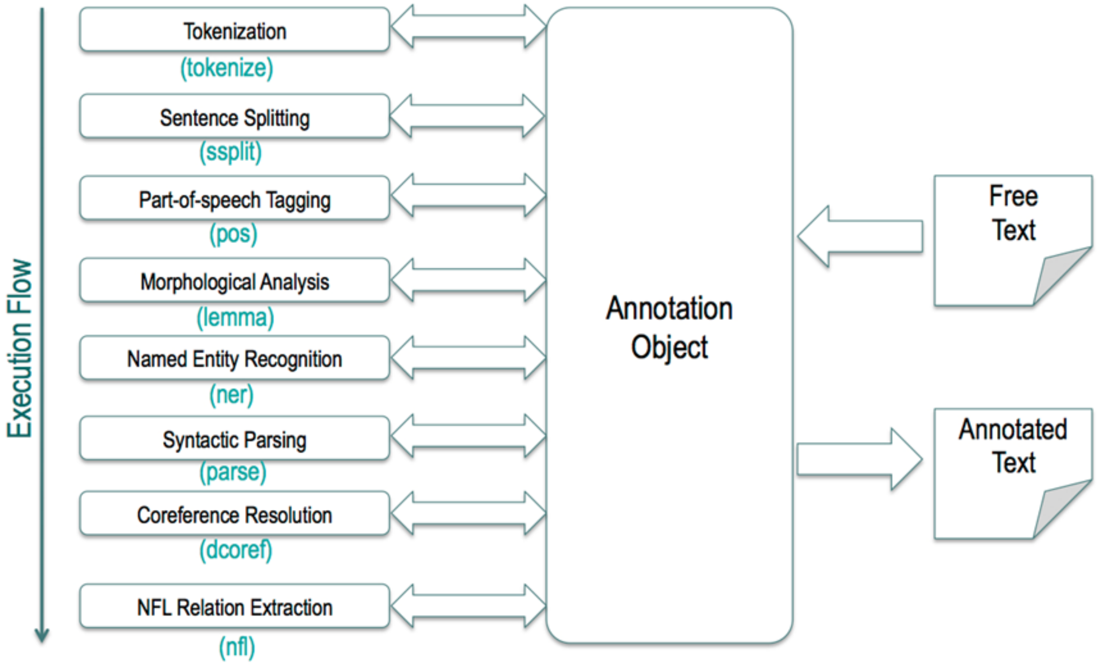


Figure 4: The Stanford Core NLP pipeline as used by the FAUST team in Phase 2 evaluation

We now provide a executive summary for each of our team members that covers the FAUST project in its entirety.

4.2.1 Stanford University (Prof. Manning)

The Machine Reading program provided a research direction and setting for Stanford to make major advances and explore new directions in natural language understanding, at levels ranging from providing general infrastructure components useful to many groups to cutting-edge research into new models of language. At the spectrum's practical end, a key result of the Machine Reading program was the development of Stanford CoreNLP, a simple-but-flexible pipeline framework that ties together all of Stanford's core NLP components, from sentence splitting and tokenization through parts-of-speech, named entities, to parsing and co-reference, and makes them available under a simple uniform API. Part or all of CoreNLP was used by most of the groups in the FAUST consortium, including SRI, UW, Wisconsin, UIUC, and UMass, and it supported the relation-extraction evaluation. However, beyond this, CoreNLP was made publicly available open source, and it has been used by many other groups, including being one of the processors used for creating the recently released LDC Annotated English Gigaword corpus produced at Johns Hopkins.

Stanford also took advantage of the problems and needs arising in the Machine Reading program to develop and release new state-of-the-art and practical NLP components for several tasks. Stanford developed a new deterministic sieve architecture for entity co-reference based on the idea of making easy decisions first and then using the emerging entity clusters to guide later decisions. This system was the best performing system at the CoNLL 2011 Shared Task [68] on entity co-reference (organized by BBN using GALE OntoNotes data), and was variously used or was the conceptual basis of three of the four best performing systems in the CoNLL 2012 Shared Task on multilingual co-reference resolution. Stanford also produced new systems for matching patterns over token sequences and dependency graphs and developed a state-of-the-art system for the recognition and interpretation of temporal mentions (SUTime).

A major focus of the work in machine reading at Stanford was the development of relation-extraction systems (finding semantic predicates and their arguments). This was explored using both fully supervised methods over linguistic analyses, as in the Phase 2 evaluation on NFL game reports, and more extensively by considering the task of distantly supervised learning, where you have some texts and an initial knowledgebase that you wish to extend with more texts. The relation between the initial knowledgebase and text gives you some guesses as to how relations are expressed in language, but that knowledge is uncertain and noisy. Stanford worked on this problem extensively in the context of the NIST TAC KBP task, and developed a new, principled model that handles the uncertainties of distantly supervised learning, the Multi-Instance, Multi-Label Relation Extraction (MIML-RE) model.

Pushing the frontiers of research, Stanford concentrated in three main areas: Stanford explored the usage of joint learning methods within NLP, doing things such as showing gains from doing joint named entity recognition and parsing, or doing successful joint learning over texts from different domains and genres. Stanford focused on extending work on co-reference and event extraction. In particular, a new model of cross-document joint entity and event co-reference was produced. Stanford was able to show that information about event co-reference aided decisions on entity co-reference and vice-versa. Finally, Stanford initiated a major exploration of deep learning (multi-layer neural network) methods for use of the data-dependent recursive hierarchical structures of natural language. This lead to the development of several new models for handling composition within vector spaces, NLP applications to parsing, sentiment analysis and relation classification, and the application of these methods to both vision and language, which won an ICML 2011 best paper award [70].

4.2.2 University of Wisconsin Madison and Wake Forest

The contribution of the University of Wisconsin and Wake Forest University to the FAUST projects was four-fold:

(1) Wisconsin contributed to the design, development, and integration of various modules in FAUST. These components included modules to perform very-large-scale joint inference and learning for and from reading. Wisconsin also contributed to several aspects of the feature engineering and the development of domain background, which proved to be essential to successful performance on the TAC-KBP and MR-KBP tasks. The modules that the Wisconsin team developed for the KBP tasks are Tuffy (a Markov Logic network RDBMS-based inference engine, which has been downloaded more than 5000 times) [48,84] and Felix (an operator-based relational optimizer for statistical inference) [85]. Wisconsin also developed and integrated an end-to-end system for MR-KBP, which takes as input a set of MR queries and a corpus of text, and then outputs extracted assertions that can be directly evaluated.

(2) Wisconsin and Wake Forest collaborated extensively to develop novel approaches and algorithms to address several open problems in Statistical Relational Learning (SRL). These approaches were quite effective when applied to MR and other text-based datasets. These included RDN-Boost and MLN-Boost [93]; functional-gradient boosting algorithms for relational dependency networks (RDNs); and Markov Logic networks (MLNs), respectively. Wisconsin also developed and implemented new approaches for very-large-scale inference, including optimization approaches such as dual decomposition and partitioning-based inference algorithms. Wake Forest also collaborated extensively with SRI to develop and test the Anytime Lifted Belief Propagation (ALBP) algorithm.

(3) Wisconsin extensively tested and successfully applied their implementations to various datasets such as NFL; TempEval-2010; TAC-KBP; and MR-KBP, working on the entity-linking and slot-filling tasks. Wisconsin developed several auxiliary tools to integrate their algorithms, including tools for cross validation, example creation, debugging, and visualization. Of particular significance is that Wisconsin demonstrated their approach of using probabilistic logic to extract information from text scaled to a corpus of more than one billion documents.

(4) Wisconsin and Wake Forest published their design, analysis, and findings of the various technologies and research developments in several high-quality conference proceedings and journals, disseminating their work to a large audience in the wider ML/AI/NLP/DB communities. Wisconsin's novel algorithms led to open-source software (Tuffy, Felix, and RDN-Boost) and publicly available demos (DeepDive and Wisci).

4.2.3 University of Washington (UW)

Prof. Weld, Task: Learning from Reading: The University of Washington pioneered and extended a diverse set of approaches for distant supervision of relational extractors. Their methods used background knowledgebases ranging from Wikipedia, Freebase, and the Nell KB, and matched to a variety of textual corpora including Wikipedia and newswire text. Their LUCHS system generated extractors for more than 5000 distinct relations [21], which is several orders of magnitude more than previous systems. Their MultiR system included a novel graphical model that not only relaxes the common prior assumptions of disjoint relation tuples, but requires two orders of magnitude less computational time than previous multi-instance methods. Finally, their VELVET system introduced the notion of ontological smoothing, a method for quickly training a relational extractor with only a handful of positive examples.

Prof. Domingos, Task: Joint Inference: The University of Washington worked toward an end-to-end solution to machine reading that builds on top of unsupervised semantic parsing and enables efficient large-scale JI. The main thesis of FAUST is that machine reading can be achieved through massive JI. However, the current JI methods were not robust enough to perform on the scale required by the project. Team goals were three-fold: (1) to create new architectures and algorithms for efficient, large-scale probabilistic joint inference; (2) to develop algorithms that unify probabilistic and logical inference; and (3) to develop methods for scalable semantic parsing from text. Integrating these algorithms into FAUST would enable fact- and rule-extraction from text and JIs based on the extracted information.

The University of Washington completed work toward these goals over all phases of the Machine Reading program. In brief, UW developed (1) USP, an algorithm for unsupervised semantic parsing, taking steps toward making it online and more scalable; (2) the CFPI framework for coarse-to-fine probabilistic inference; (3) PTP, a new approach for unifying logical and probabilistic inference; (4) ABQ, a new approach for efficiently conducting approximate probabilistic inference; (5) SPNs, a new deep architecture that is more general than arithmetic circuits and also enables efficient exact inference; (6) a theory of USPN, an end-to-end solution to machine reading that would extend USP to process text online; (7) a family of deterministic, structured message-passing algorithms for efficient JI; (8) an algorithm for multiple hierarchical relational clustering; (9) TML, a tractable subset of Markov Logic that can be used for logical-probabilistic representation and tractable JI over the entire machine-reading process, including syntactic and semantic parsing, ontology and knowledgebase population, and question answering; and (10) a linear-time shift-reduce CCG semantic parser. Several papers described these results [1,43,72,73,74,78,84].

4.2.4 University of Massachusetts Amherst

The University of Massachusetts Amherst contributed along several dimensions of the Machine Reading spectrum: joint models, new learning and inference algorithms, and software libraries.

Models: UMass developed a joint model for event extraction that combined entity-type prediction and detection of event arguments. Inference was done using dual decomposition. This model ranked first in the BioNLP shared task. UMass built the first cross-document joint Named Entity Recognizer (NER) and relation-extraction model, trained only with weak supervision.

Learning: UMass developed SampleRank, a highly scalable algorithm for learning in large-scale graphical models. This algorithm supports arbitrary, user-specified loss functions, and trains models both more quickly and more accurately than previous methods. UMass pioneered a new paradigm for distant supervision by introducing latent variables that indicate whether a relation is expressed by a mention. This paradign has improved the accuracy of relation extraction, and has already sparked a long line of follow-up work, including contributions from other FAUST members.

Inference: Joint Inference was the core theme of FAUST, and UMass developed several algorithms to make JI scalable. These algorithms were orders of magnitude faster than previous methods. The speed was achieved by lazily instantiating both factors and variable values only when they were needed. UMass developed a new generation of cross-document co-reference algorithms that rely on hierarchies of co-reference clusters for both increased robustness and efficient parallel inference.

Software: UMass developed the FACTORIE toolkit for deployable probabilistic modeling, implemented as a software library in Scala. It provides its users with a succinct language for creating relational factor graphs, estimating parameters, and performing inference. UMass also released their BioNLP event extraction and contributed the IC domain information-extraction component of FAUST.

4.2.5 University of Illinois Urbana-Champaign (UIUC)

Prof. Roth: The team headed by Dan Roth worked on several tasks, contributing to the efforts on Joint Inference, Natural Language Processing (NLP), and Learning for and from Reading. UIUC pioneered a framework to support incorporating declarative knowledge as a way to guide learning and support global inference. The Integer Linear Programming-based formulation has been used and was the subject of research by other team members including UMass and Columbia. In particular, UIUC developed new algorithms for learning with indirect supervision, and for learning and inference with latent representations. This framework was used in developing multiple NLP capabilities, including (1) relation and event extraction; (2) co-reference; (3) textual inference; and (4) temporal and causal reasoning. UIUC's key contribution to the learning task was the Wikifier, an approach for disambiguating concepts and entities appearing in text and grounding them in an encyclopedic resource. This is both a knowledge-acquisition tool and a way to support co-reference within and across documents and other textual inferences.

UIUC continued developing better NLP analysis tools throughout the project and made them available to the research community. UIUC made available the Curator, a distributed system for running and aligning multiple-state NLP preprocessing tools, as well as state-of-the-art tools for multiple NLP tasks, including semantic role labeling, named entity recognition, and co-reference resolution. SRI and several of the other team members used these tools.

Prof. Amir: In Phase 3R, Prof. Amir joined the FAUST team. This team worked on probabilistic modal (PM) operators for natural language understanding. PM models were investigated to represent what authors assume about readers' knowledge. Both (1) a theoretical framework for inferring Bayesian Network PM models from text and (2) an implementation of that framework in computer algorithms and executable programs were created. PM models were extended to dynamic domains in which actions change the state of the world. These models capture events in NL texts and enable modeling the beliefs of authors about beliefs of readers about those events and their participants. The effects of actions are modeled as stochastic choice between deterministic executions.

4.2.6 PARC and Stanford's CSLI

PARC and CSLI's work focused on inferences that can be drawn from texts based on inferential properties of linguistic expressions. Such inferences are a necessary part of automated NL understanding. This work demonstrated that the task can be aided by different kinds of resources, including lexical class markings, ontological classifications, and domain models that link different classes of items together.

CSLI based their study of the veridicality inferences of texts on the following broad hypothesis: (1) a large class of lexical items in particular syntactic frames, or specific types of phrases, are associated with a veridicality signature; (2) the implications of whole sentences about their author's commitments arise from a projection mechanism from the veridicality signatures of the elements embedded in them; and (3) contextual factors might strengthen these implications. This work then involved the following three components:

- Analysis to determine the range of veridicality signatures and the environments in which lexical items have them and to identify and understand the contextual factors involved
- Verifying the analysis with human subjects
- Figuring out an algorithm of projection and effect of contextual factors

4.2.7 SRI International Research Team

SRI's research team developed *Lifted probabilistic inference,* which manipulates the representation in first-order form, keeping it compact and performing operations on a single conditional probability function. By contrast, regular inference would perform the same operations repeatedly for each instance of that function, requiring exponential effort. SRI developed an engine for *anytime lifted probabilistic inference.* If the query depends only on a small fraction of the model, as most do, then the algorithm will not examine the entire model.

SRI developed a formal notation and representation to describe such algorithms without ambiguity. Because this representation enables casting lifted inference as a form of *symbolic evaluation*. SRI developed a Lifted Belief Propagation (LBP) algorithm, implemented as symbolic evaluation. SRI released the software of the probabilistic inference engine, the symbolic evaluation system, and general utilities, as three separate projects.

SRI developed a new lifted inference method, LIDE (Lifted Inference with Distinct Evidence), that allows polynomial-time exact lifted inference even in the presence of unique evidence on a set of grounding instances of a unary predicate, one for each individual [138].

SRI, in collaboration with UMass, developed a general framework for lifting variational approximation algorithms [150] such as linear programming relaxation of maximum *a posteriori* (MAP) inference, a widely used approximation in NLP problems. Initial experimental results demonstrate that lifted MAP inference with cycle constraints achieved state-of-the-art performance, obtained much better objective function values than local approximation while remaining relatively efficient (order-of-magnitude faster than inference on the ground model).

4.2.8 MIT and Columbia University

Inference: JI was a core focus of work in Phases 1 and 2, initially at MIT, which then moved to Columbia University. Columbia developed novel methods based on dual decomposition and Lagrangian relaxation for inference in NLP. In this approach, constraints that make a problem computationally challenging (e.g., NP hard) are relaxed, through the introduction of Lagrange multipliers. A subgradient algorithm is used to minimize the resulting dual. Various guarantees can be derived, including a guarantee of optimality if the algorithm converges to a point where the constraints are satisfied when decoding under the penalized primal problem. The method was shown to be effective in a number of NLP problems. The work resulted in several publications (including a best paper award at EMNLP 2010) [36], and a tutorial at ACL 2011[1].

Learning: work at Columbia in Phases 3 and 3R focused on the development of spectral-learning methods for latent-variable models. The EM algorithm is a widely applied method in NLP. However, it is well known to only give locally optimal solutions. Recent work has introduced spectral methods as an alternative to the EM algorithm for learning in latent-variable models. Novel spectral-learning algorithms for latent-variable PCFGs were develop. Recently completed experiments show that these methods perform at the same level of accuracy as EM, but are an order of magnitude more efficient in training time.

[1] http://www.cs.columbia.edu/%7Emcollins/papers/dual_decomp_tutorial.pdf, Dual Decomposition for Natural Language Processing, by Alexander M. Rush and Michael Collins.

4.2.9 Institute for Human and Machine Cognition (IHMC)

IHMC joined the team in Phase 3R and executed an exploratory project to locate proto-beliefs of individual Ummah message board posters on a large scale. These beliefs could then be examined to determine the consistency of an individual poster's beliefs and to identify where that individual's beliefs conflict with the beliefs of others; such conflicts of belief could occur either within the context of a single thread or in the context of all threads.

In the information flow of the completed system, facts were extracted from the Ummah message board postings using unsupervised methods for information extraction. These facts were then linked to individual posters as beliefs or assertions in a belief management engine. Finally, heuristics were used to investigate confirmations and negations of beliefs within and outside individual message threads.

An exploratory effort was pursued to determine the feasibility of this approach to the extraction and comprehension of agents' interrelated beliefs. The primary outcome of the completed work is a positive demonstration of the extraction of these beliefs. In particular, it was demonstrated that beliefs could be (1) extracted from the unstructured data contained in an online forum, (2) represented in the ViewGen belief engine, and (3) scored using heuristic approaches similar to the FactRank (Jain & Pantel, 2010) algorithm.

4.2.10 Onyx Consulting

Onyx joined the team in Phase 3R and studied elided scopes of modality. Ellipsis is a linguistic process that renders certain aspects of text meaning invisible at the surface structure, thereby making them inaccessible to most current text-processing methods. Ellipsis is considered one of the more difficult aspects of text processing and, accordingly, has not been widely pursued in NLP applications.[2] However, not all cases of ellipsis are created equal: some can be detected and resolved with high confidence within the current state of the art. Onyx has been working toward configuring a system that can resolve one class of elliptical phenomena: elided scopes of modality. Onyx addressed the problem of elided scopes of modality from two perspectives:

1. Onyx developed a full microtheory of modal-scope ellipsis treatment that is being incorporated into the language-enabled intelligent agents in the OntoAgent cognitive architecture. This direction of work is reported in the conference paper "Resolving Elided Scopes of Modality in OntoAgent" [151], which was presented at the First Annual Conference on Advances in Cognitive Systems (December, 2012. This approach employs all of the static knowledge resources and reasoning engines available to OntoAgent intelligent agents.

2. Onyx developed a method of detecting and resolving a subset of cases of modal scope ellipsis that can be applied to big data. To work over big data in real time, the approach uses only a subset of the resources and reasoners available in this environment and replaces some of the more resource-intensive aspects of processing with cheaper proxies. The goal was to focus on achieving high precision over a large dataset.

[2] As Spenader & Hendriks (2005) write in the introduction to the proceedings of a workshop devoted to ellipsis in NLP, "The area of ellipsis resolution and generation has long been neglected in work on natural language processing, and there are few examples of working systems or computational algorithms." In fact, of the ten contributions to that workshop, only one reports an implemented system, the others discussing corpus studies of ellipsis, descriptive analyses of phenomena, or theoretical (typically, pragmatic) frameworks in which ellipsis might be treated.

4.3 PHASE 1 RESULTS

The main technical foci for Phase 1 were (1) defining the first version of the FAUST architecture and setting up the software-engineering and software-management infrastructure for the project at SRI; (2) developing the machine-learning software that will execute the Phase 1 Readability Assessment evaluation; and (3) initiating a large number of explorations of various joint-inference regimes, both among NLP components and between such components and the main components of the envisaged overall Joint Inference capability. With respect to (2), we built a number of multi-feature Readability classifiers, some of which included a number of novel discourse-level features aimed at capturing degree of discourse coherence. We performed initial tests of the classifier and made adjustments, and we took and passed the Phase 1 Readability Assessment Evaluation.

The motivation for (3) was to explore a significant part of the broad space of possibilities for implementing the FAUST vision of a reading system based on joint inference across the full range of capabilities required for Machine Reading. Below, we report on a number of such explorations, many of which involve multiple institutions.

We report our results organized by the major tasks. Multiple institutions contribute to each task, and some institutions contribute to multiple tasks.

4.3.1 Natural Language Processing

4.3.1.1 Stanford

Stanford completed the implementation of a baseline supervised model for the extraction of entities involved in relations of interest (as defined in the MR evaluation domains or for other relation-extraction tasks). This relation-extraction system is a core component of several pieces of this machine reading work. The supervised model works in two steps: (1) it extracts the syntactic head of the annotated constituent for each entity mention (e.g., “rout” for the phrase “a 44 to 15 American football rout of Chicago”) and (2) it classifies these mentions into corresponding classes using a linear-chain CRF (e.g., the previous mention is classified into the NFLGame class). This system implements lexical, syntactic, and gazetteer-based features.

- In the NFL domain, the Stanford system achieved an overall F1 of 70 across different entity types. For detecting relations, it achieved an F1 score of 72 when using gold entity mentions on the current version of the NFL domain corpus, but an F1 of only 23 when using predicted entity mentions. In December, Stanford looked at the BioNLP domain, where an F1 score of 92 for detecting protein entities was obtained.
- Stanford also worked on incorporating new corpora and domains, such as newswire and ACE, and defined a unifying representation for relations and events that covers all these domains. Stanford aimed to evaluate their system on corpora other than the NFL corpus because (1) it was too small for a relevant analysis and (2) it has limited usefulness for tasks important to MR. In addition to doing entity detection, Stanford worked on extending their system to do relation and event extraction. This work was to accommodate Use Cases 3 to 6.

Stanford participated in NIST TAC KBP 2009 (Knowledgebase Population). This was the very first run of this task, a “machine reading” task with considerable overlap with some of the knowledge-extraction tasks that were major goals of the MR program.

As part of TAC KBP entity linking, Stanford build an initial system for entity-linking and composed of a dictionary mapping strings of text to the potential Wikipedia pages that they can refer to [39]. The dictionary is built using the anchor text and links connecting Wikipedia pages together, and it includes frequency statistics on how often a string is used to link to a particular Wikipedia page.

With an eye to TAC KBP slot-filling and other relation-extraction tasks, Stanford started an investigation on the effect of distant supervision on an existing supervised Information Extraction (IE) system. While previous work had shown that a system trained from distantly supervised corpora performs well when evaluated in the same environment, it was unclear that adding a corpus generated through distant supervision to an existing supervised system would improve performance. Stanford aimed to answer the following questions:

(1) Does distant supervision improve the quality of a supervised system?

(2) Does Amazon's Mechanical Turk (AMT) help improve the quality of a corpus generated through distant supervision?

(3) What is more important for distant supervision: quality (i.e., validating data through AMT) or quantity (i.e., automatically acquiring large amounts of data)?

To answer these questions, Stanford built a framework that enables combining distantly supervised and supervised approaches [38]. This involves extracting distantly supervised relation instances from Freebase and getting sample data by mapping these instances to Wikipedia sentences. These sentences are optionally validated by a set of Amazon Mechanical Turk annotators. This framework supported both the NFL domain (first Machine Reading IE evaluation) and Intelligence Community (IC) slot-filling tasks (part of IC Use Cases 3–6).

Stanford built a new deterministic co-reference resolution system. The co-reference system focuses on the deployment of a large set of features, ranging from agreement to syntactic and semantic constraints. When the system was approximately 75% complete, it already outperformed Stanford’s then current co-reference resolution system. On MUC, it achieved a pairwise F1 score of 62.8. The state-of-the-art system (Haghighi and Klein) for this dataset was 67.3, but this included additional features not present in Stanford’s model.

In a related project, Stanford investigated whether using additional non-expert annotations from Amazon’s Mechanical Turk can improve the training of supervised classifiers. Stanford used the NFL corpus as the base dataset in a 10-fold cross validation and tried different approaches to relation and entity extraction. In order to generate an additional dataset, Stanford first crawled the web extracting sentences that have the same type of features as the original annotated NFL set. Then Stanford used all appropriate entity combinations to generate a large set of relation candidates and then finally used Amazon’s Mechanical Turk to prune the sets of relations for a given sentence. By pruning relations, users also prune entity types. The results show that this improves the accuracy of the entity extractor. On the other hand, even though the size of the training set significantly increased, the relation extraction did not perform better and in fact had lower scores for several relation types. There are several observations that one can make based on the results. First, when using additional non-expert annotations, using a classifier that offers the capability to work with various degrees of confidence for annotations is important. In the case of an existing relation classifier, every relation that is not marked explicitly as positive is turned into a negative relation. This significantly hurts the performance of relations with fewer annotations. Another result is that the quality of annotation has significant impact on the subsequent performance of classifier. Thus, another important aspect is using more sophisticated approaches for selecting annotations than simple majority voting.

Stanford created an unsupervised system for ranking events by their durations. The system is able to provide both coarser duration classification (by assigning events into the duration buckets: seconds, minutes, hours, days, weeks, months, years, and decades) as well as more fine-grained ordering within each bucket. Overall, the system is able to rank the list of events from the shortest to longest without use of any supervised training data. This approach uses a set of web queries to create distributions of hits across duration buckets and then uses several ranking algorithms to predict the duration of an event based on the given distribution. The only input into the system is cardinality of hit result sets for each set of queries.

Stanford built a high-accuracy, fast, linear-time, semi-supervised dependency parser.

- Stanford performed an initial study on ensemble models for parsing [10]. This study yielded several observations relevant to MR: an ensemble of models of linear-time complexity (in the number of tokens in a sentence) outperforms existing state-of-the-art parsers that require polynomial time to parse a sentence. Ensemble models that combine both linear-time and cubic-time models achieve comparable performance to the best models in the world for the parsing of syntactic dependencies, which have much higher overhead. This study indicates that fast parsing of large-scale corpora is possible without any relevant loss in parsing accuracy.
- By creating parsers using non-linear kernels (polynomial with degree 2), accuracy was substantially improved and creating an ensemble of these models improved results further. However, the models with polynomial kernels do not operate in linear time. Stanford hoped the parsers with polynomial kernels could self-train the simpler and faster linear-kernel parsing models. Experiments explored using both single parsing models and ensemble models to do the self-training. In this phase, Stanford did not find a way to make this method improve over the baseline. That is, while individual linear-kernel parsing models are actually improved when self-training from an ensemble of polynomial kernel parsing models, an ensemble of the linear kernel parsing models does not.

4.3.1.3 PARC

PARC built and delivered to SRI LexBase, a lexical database manager. It reads the terms and the lexical and semantic relations defined by WordNet (see http://wordnet.princeton.edu) and then stores them in a memory-resident database. LexBase enables querying of this database, looking up nouns, verbs, adjectives, and adverbs, and retrieving related words and concepts, such as synonyms, antonyms, hypernyms, meronyms, and so on. Beyond replicating the functionality of WordNet, LexBase supports programmatic editing of the database. There are commands to add or delete terms, word senses, and relationships from the database.

PARC built BD-1, an indexing system that provides a scalable, flexible database for retrieving complex linguistic structures. BD-1 is a database system for storing and querying of n-tuples. The system is designed specifically to provide efficient search and natural representations of annotated text. These annotations can consist of text fragments or (for look-aside annotations) may reference text spans; the values stored in BD-1 can be string or binary data (e.g., integer values). As a generic database, BD-1 can function as a key-value database, a triple store, or an n-tuple store. BD-1 is compatible with the Berkeley database and supports a query language for n-tuples that is a simplified subset of the SPARQL query language for RDF. It can be configured to use memory as a cache for its data store—which is particularly useful for lexical resources that can easily be accommodated in current machines, such as WordNet.

PARC worked on the specification of underspecified content based on implicit ontological classifications. Machine reading requires a level of natural language processing that enables direct inferences to be drawn from the processed texts. Although most heavy-duty inferencing has to be done by a reasoning engine working on the output of the linguistic analysis (with possible loops between the two), the linguistic analysis should deliver representations where a certain level of disambiguation and content specification has been done. The pervasive ambiguity of language enables sentences that differ in just one lexical item to have rather different inference patterns. An illustration of the problem comes from sentences of the form "A went from X to Y," which can be used to describe movement or spatial extent of an entity, or the change in values of a fluent across time or space, depending on the ontological properties A, X, and Y are assumed to have. For example, "Thacker went from PARC to Microsoft Research" can imply a change in Thacker's location or a change in Thacker's employment. The problem is that different lexical items do not fall into clearly definable and easy to represent classes. To draw the correct inferences, one needs to look how the referents of the lexical items in the sentence (or some broader context) interact in the described situation. In order to perform an interpretation without a model of the domain, however, one needs to find the features of words that can be used to provide appropriate guidance. In collaboration with the University of Washington, PARC conducted a corpus analysis to find characterizations of appropriate features. Moreover, PARC's work on "from/to" phrases showed that as modifiers of the main predication in a clause, they introduce paths for extent, change, or scales, and can have both an independent and a correlated interpretation. The latter express a functional dependence between the two paths, as in "The temperature went from 50 to 90 degrees F from the top of the mountain to the bottom."

PARC's temporal expression annotation technology was used to annotate the IC MR corpus. PARC formed a working group with members of Stanford's Natural Language group to explore how to integrate interpretation of temporal modifiers into Stanford's system.

PARC explored the inferential properties of different clause types. One particular issue is the following: if an assertion of ‘p’ is used to commit the speaker to the belief ‘p’, what type of commitment does an utterance of an imperative ‘p!’ give rise to? Taking as a starting point the idea that imperatives express commitments to act as though a certain agent had a certain preference, two intertwined questions were investigated: (1) Whose commitment to preferences do imperatives talk about? (2) What kind of preferences do imperatives talk about and how can they be represented formally? PARC introduced the notion of a preference structure, which serves as a general tool for encoding (ranked) preferences of agents.

4.3.1.4 UIUC and UMass

Several UIUC NLP software packages were committed to the SRI repository. UIUC re-factored some of the code to use a uniform name space per SRI’s request. Key changes were in the named-entity recognition, named-entity similarity, and co-reference packages.

- UIUC developed extensions to the existing semantic-role labeling package, with the goal of extending SRL beyond verb predicates to a number of other relations.
- UIUC improved named-entity resolution and co-reference resolution. The emphasis is on co-reference across documents and on Wikification—mapping entities and concepts to Wikipedia.
- UIUC worked on a better mention-detection approach, to aid both in co-reference, named entities and Wikification; the emphasis is on incorporating this module in other tools in a modular way, without affecting the trained model, to support moving to a new domain.
- UIUC worked on learning to identify generic relations such as “is-a” and “sibling” between concepts.
- UIUC worked on relation recognition and on event recognition, tracking, and de-duplication.
- UIUC worked on supporting integration of multiple levels of natural language analysis. This included a new multiview-based alignment algorithm, which incorporates multiple levels of analysis of natural language—including POS; shallow parsing; dependency parsing; semantic role labeling; named entities; and co-reference resolution as a way to align text and hypothesis in the context of textual entailment [46].

UMass made progress in joint morphology extraction and parsing with Belief Propagation. Inference was made faster by exploiting substructure in large factor graphs and sparse distributions from morphological dictionaries.

UMass continued work on large-scale, cross-document co-reference with distant supervision. UMass demonstrated performing co-reference on five million mentions from the *New York Times* using Wikipedia as distant labeling and employed a CRF-based co-reference model to attain approximately 90% F1. UMass began developing a new, more scalable and more expressive model based on representations of co-reference decisions in a hierarchy.

UMass continued to develop a state-of-the-art co-reference system in FACTORIE [4], which is designed to support joint inference with various other tasks. UMass enhanced co-reference infrastructure and features, including the addition of latent canonical entity variables, and Haghighi-style latent cross-document multinomials.

UMass continued work on joint syntactic and semantic parsing and learning to generalize semantic frames.

4.3.2 Joint Inference

4.3.2.1 SRI and Wisconsin

SRI worked, with Wisconsin collaboration, on the design and implementation of the anytime lifted belief propagation (ALBP) algorithm. This algorithm represents a significant step in efficient and scalable inference. It addresses a significant limitation in existing approaches to lifted inference wherein a model is shattered (variables divided into manageable clusters) *before* inference. This approach performs shattering *during* inference at the expense of trading-off exactness of belief of the queries with a bound, or range, on the beliefs. This *anytime belief propagation* interleaves shattering and inference and is able to obtain exact bounds on the query.

Wisconsin developed a Java-based, large-scale inference engine called Tuffy. It leverages the full power of a relational optimizer in an RDBMS to perform the grounding of MLN models several orders of magnitude faster than the current state-of-the art. While MLNs represent a powerful formalism for inference, existing approaches to MLN inference did not scale to larger real-world datasets. Tuffy performs maximum *a posteriori* (MAP) inference and achieves orders-of-magnitude improvement in scalability, compared to existing MLN approaches through three novel contributions:

- *Bottom-up Grounding*: This is in sharp contrast to existing approaches, which perform top-down grounding and use inefficient techniques such as nested loops. Tuffy expresses grounding as a sequence of SQL queries, each of which is optimized by the RDBMS (Tuffy uses PostgreSQL as the default relational database), resulting in a great speed-up in grounding.
- *Hybrid Architecture*: While performing an AI-style search within the RDBMS is possible, Tuffy employs a hybrid architecture, where the inference is performed in-memory, which is more efficient. This is the case because the ground MLN is a Markov random field (MRF), and inference essentially reduces to a Boolean satisfiability problem that can be most efficiently solved locally and in-memory.
- *Partitioning to Improve Performance*: The time and space efficiency of Tuffy was also improved by decomposing the problem into smaller pieces with minimum information loss. This enabled Wisconsin to use component-aware and partition-aware search algorithms *in parallel* on each partition, resulting in further gains in scalability.

Tuffy is open-source and can be found at http://hazy.cs.wisc.edu/hazy/tuffy/. For the architecture and further technical details, see [48][3].

[3] http://hazy.cs.wisc.edu/hazy/papers/tuffy-vldb11.pdf

4.3.2.2 Other Team Members

The University of Washington's Phase 1 objective centered around developing and implementing an initial version of an algorithm for efficient lifted inference utilizing a hierarchy of types. Current joint inference (JI) methods are not robust enough to perform on the scale required by FAUST. While newer lifted-inference methods give some gain in efficiency, the shattering process that creates the lifted networks can be a huge bottleneck, and lifted inference may still be infeasible on very large datasets. Meeting this objective enables running the large-scale JI required by FAUST.

To this end, UW developed Ontological Belief Propagation (OBP). OBP is an approximate-inference algorithm that runs lifted belief propagation in an iterative, coarse-to-fine manner. At each stage in the procedure, low-probability areas are pruned from the search space for the next, more refined stage. The main efficiency bottleneck in lifted-inference algorithms is the initial shattering procedure that creates the lifted network on which inference is then run. OBP removes this obstacle by taking advantage of a hierarchy of types to shatter at different levels of refinement.

UW implemented and evaluated the Ontological Belief Propagation algorithm. Experiments were performed on a link prediction task in a social networking domain and a biomolecular event prediction task. For the link prediction task, an order of magnitude speedup over lifted belief propagation with virtually no change in accuracy was obtained; the algorithm also showed impressive efficiency gains in the event prediction task without loss of accuracy. Error bounds of both Ontological Belief Propagation and the more general Ontological Lifted Probabilistic Inference (OLPI) framework were studied. Final results were presented in an AAAI-11 paper [78], in which the name of this framework was changed to Coarse-to-Fine Probabilistic Inference (CFPI) to generalize it beyond situations involving ontologies.

UMass further developed and evaluated "relaxed" marginal inference, which carefully selects which factors may be safely ignored (the resulting method is loosely related to "cutting plane" methods, but designed to operate on marginal inference instead of MAP inference optimization). An evaluation was performed on dependency parsing with belief propagation, leading to 20-fold speedups and the reduction of graphical model size without a loss in accuracy (and even sometimes surprisingly seeing gains in accuracy). This methodology will support parsing with extra arbitrary dependency structures not possible with traditional dynamic programming, which ultimately could reach all the way into the KB. The NAACL paper [12] described the better second-order models that lead to an additional 1% absolute improvement (and best results for the CoNLL 06 Dutch dataset). A paper on the methodological aspects of this work, with further evaluation, was published in UAI [17].

UMass further developed infrastructure for joint co-reference and relation extraction. As part of an effort toward substantial testing and evaluation, they began work on the WEPS competition (web people search) using FACTORIE, including the design and implementation of a simple, baseline system architecture that combines co-reference and relation extraction in a pipeline.

UMass began updating the publicly released version of FACTORIE to work with the latest version of the Scala compiler to better support probabilistic database integration.

UMass investigated scalable probabilistic database representation with FACTORIE using object-oriented database technology, after finding that db4o did not scale. Working with BerkeleyDB

“Java Edition” and Terracotta yielded better results. An added advantage of Terracotta is that computation can be distributable across many machines.

UMass demonstrated an approach to arbitrary factor graphs in probabilistic databases backed by a plain relational SQL database (MySQL). This was demonstrated on entities from 10 years of *NY Times* articles.

Stanford created a method for improving joint models using additional data that has not been labeled with the entire joint structure. This was done by: building single-task models for the non-jointly labeled data; designing those single-task models so that they have features in common with the joint model; and then linking all of the different single-task and joint models via a hierarchical prior. The results on joint parsing and named entity recognition showed that the new model substantially outperformed a joint model that was trained on only the jointly annotated data by an absolute f-score of 8% on both parsing and named entity recognition.

MIT worked on developing algorithms for JI based on linear programming (LP) relaxations. LP relaxations have recently been applied with great success to inference in Markov random fields (MRFs). MIT developed results that connect dynamic programming algorithms used in NLP (e.g., for parsing) to linear programming problems. This provides a theoretical underpinning for the use of LP relaxations. Building on this, dual decomposition methods for parsing and other NLP problems were developed [36,37]. The approach enables JI across two or more models to be achieved using a very simple algorithm based on black-box solvers for the two models, combined with a sub-gradient descent on Lagrange multipliers that enforce agreement between the different models. As a proof of concept, a second-order discriminative dependency parser was combined with a generative constituency-parsing model to achieve competitive results (and significant improvements over a naïve combination method that uses the dependency structures as hard constraints on the generative model).

4.3.3 Learning for and from Reading

Wisconsin developed an approach for using Inductive Logic Programming (ILP) plus MLNs to learn patterns from the sample extractions provided for a domain. This approach performs pre-processing to substantially reduce the size of the ground Markov Logic network. This is done by counting how often the evidence satisfies each formula, and does not consider the truth values of the query literals; it does however only consider a very small fraction of all possible groundings. This results in an algorithm called Fast Reduction of Grounded Networks (Shavlik et. al., 2009).

Wisconsin developed, encoded, and tested extensive background knowledge for the NFL test bed. The goal was to investigate what types of background knowledge were needed to understand the sentences about NFL games. Wisconsin used:

- General knowledge about English (from the Stanford NLP Toolkit)
- Knowledge about temporal statements (e.g., “this week” and “next month”)
- General NFL knowledge, such as team names and nicknames.

Wisconsin developed some “background theories” or background concepts, especially for temporal relations (e.g., how to convert “this Friday” or “next month” into an absolute date given the specific date an article was written), and scores in sporting events (e.g., recognizing that “14-7” is a possible NFL score but that “81-80” likely is not). Wisconsin specified essential prior knowledge/expert advice (e.g., “there is only one winner in a game,” “a touchdown is worth 7 points”). The Wisconsin Inductive Logic Learner (WILL) can use this domain background and advice.

Wisconsin advanced its research in learning complex distributions using MLNs. Their novel approach can compile causal independencies (the notion that there can be multiple independent causes for a target variable) into MLNs. Combining rules are associative and commutative operators that combine distributions due to multiple instantiations of different rules (e.g., average-based, which average distributions over instantiations). Exploiting the fact that combining rules can capture the notion of causal independence for SRL models, Wisconsin developed an algorithm for representing a class of combining rules (called *decomposable combination functions*) in MLNs, which are an undirected model. Explicit examples provided by Wisconsin include average-based and noisy combination functions as well as a formal description of this approach that converts directed models with combining rules to MLNs. This work was described in a 2010 ECML PKDD paper [31][4].

Open Extraction: The University of Washington developed a new method for unlexicalized (open) extraction by training using distant supervision over Wikipedia infoboxes. The resulting system, WOE, demonstrated between a 73 and 107% improved F1 score compared to TextRunner. A paper on this research was published at ACL 2010 [20].

[4] http://ftp.cs.wisc.edu/machine-learning/shavlik-group/natarajan.ecml10.pdf

Distant Supervision for Relation Extraction: UW extended distant supervision, showing how to learn extractors from extremely sparse, heuristically generated, training data (which allowed them to learn 5224 different relation-specific extractors from Wikipedia—two orders of magnitude more extractors than anyone had previously attempted). The method is based on a novel lexicon-creation method, which uses the distillation of a 5 GB web crawl to create custom lexicons for each relation. A paper describing the resulting LUCHS system was published at ACL 2010 [21].

Ontology Construction: UW devised a novel method for clustering similar relations—bottom-up aggregation using a pseudo-distance function defined by training an extractor on examples of one relation and measuring the F1 score on samples of the other. Using this method, the 5224 relations encoded by the most popular Wikipedia infoboxes were clustered.

UW extended the mapping between Wikipedia and Freebase to handle properties corresponding to length-three paths. Path-selection heuristics were developed to defeat the combinatorial explosion in mapping. UW rewrote the ontology code (Weld et. al., 2008) and used a Support Vector Machine (SVM) method to create another ontology over Wikipedia infobox attributes.

Temporal Extraction: UW implemented a joint inference approach, the TIE system, to extracting facts from sentences and bounding the time interval during which they hold. A notion of temporal entropy was defined and experiments completed comparing TIE to several others. A paper describing this research was published in AAAI 2010 [18].

UIUC worked on Transfer Learning and Adaptation algorithms as a way to enable existing NLP tools to generalize better to data that is different from the training data. The current focus is de-lexicalizing NLP—attempting to learn generic modules with little domain-specific lexical information and supplying the lexical information from the outside without a need to retrain.

UIUC developed a new learning algorithm that learns a latent structured representation used as an intermediate representation for learning [7]. The current evaluation of this algorithm is done in the context of paraphrasing and textual entailment.

UIUC developed a new learning algorithm for structured prediction that can be trained using an indirect supervision signal—a signal generated in a cheap way from a companion binary decision problem associated with the structure prediction problem [14].

UIUC developed a variation of the aforementioned algorithm that enables training semantic parsers in an unsupervised way.

UMass, as part of their research on lightly supervised learning, further enhanced their new method of semi-supervised learning using a constraint-based objective function in SampleRank [12].

UMass further developed their new generative/discriminative semi-supervised learning algorithm that encourages the latent variables of rich generative models to be relevant for a discriminative task. In the previous version, constraints derived from the labeled data were used to encourage the generative model to discover relevant structure. In the new version, these constraints are iteratively refined during training. This flexibility alleviates some practical optimization issues when training with the previous approach, and preliminary results suggest that it may also provide higher accuracy [15].

UMass continued its investigation of semi-supervised learning of extractors by the alignment of text against records in a knowledgebase. They focused on the development of methodology that combines both generative and discriminative models, as well as constraints on the alignments.

UMass further improved relation extraction on the *New York Times* corpus based on distant supervision by including entity-type information in Freebase. Because this information is not available for novel entities, UMass began to implement distant supervision for both entity types and relations. Initial results achieve about 80% accuracy on entity types (without annotated data).

MIT developed dependency-parsing algorithms that enable using higher-order (trigram) dependency features, while maintaining efficient (O(n^4)) parsing algorithms.

MIT extended its work on semi-supervised learning methods. These methods learn representations from unlabeled data, which are then incorporated within a supervised approach to significantly reduce the amount of labeled data required as supervision.

4.3.4 Infrastructure, Software Engineering, and Integration

SRI’s SE team focused on ensuring that all other teams had the tools and information necessary to begin their work for the reading task. We set up our state-of-the-art software-engineering infrastructure tools and processes, including automatic builds; regression tests; issue identification; tracking and resolution; and multiple methods for enhancing quick-reaction team collaboration and reducing technical risk.

SRI began working with our team members early and assigned individual liaisons to closely collaborate with each subcontractor to define and use APIs and functional requirements and to ensure the correct use of SRI’s processes for software production.

SRI purchased hardware and set up services for source control; a wiki; issue tracking; and release distribution. We worked to interface with the MRP-provided (by SAIC) evaluation software and services (MRAPI) and integrated the software as delivered by our subcontractors in the evaluation version of our FAUST system, releasing the evaluation results.

SRI worked with subcontractors to define a common API and libraries to facilitate communication between the modules provided by different subcontractors to enable the project-wide goal of JI everywhere. This software is referred to as Common Annotation Format (CAF).

Wisconsin extended WILL to include domain (expert) knowledge. WILL is Wisconsin’s Java-based implementation for Inductive Logic Programming (ILP), which uses logic programming to represent background knowledge, training examples, and learnable hypotheses. This extension improves ILP so that it can use “advice” provided by an expert to learn stronger concepts with fewer examples. This is achieved by modifying the ILP search procedure to further take into account the expert’s advice, in addition to the training examples and background.

Wisconsin implemented MLN learning algorithms integrated them with WILL, based on fast reduction of Markov Logic Networks (MLNs).

Wisconsin developed an initial design and implementation of Tuffy, a Java-based probabilistic, deductive database management system. Tuffy performs very large-scale inference in Markov Logic Networks. The scalability of Tuffy arises from the leveraging of a relational database (PostgreSQL) to perform grounding as well as AI-style local search through novel hybrid architecture.

Wisconsin developed several modules and components for JI and Learning from Reading. These components, which were integrated into FAUST, included translator modules to convert data to WILL format and to convert annotated XML files and the output of Stanford parser to first-order logic sentences and WILL.

Stanford began work on building a usable and extendable pipeline architecture for easily accessing all of our core natural language processing tools (including named entities, parts-of-speech, parsing, and co-reference). In Phase 1, this architecture was referred to as the “Baseline Natural Language Processor,” but it was later rebranded as the “Stanford CoreNLP,” and was made generally available as open-source software in Phase 2.

Stanford worked with UW, UIUC, and SRI on integrating the Baseline NLP with their systems, and placed a copy of it in the FAUST source-code repository. Stanford developed documentation for the system and released several updated versions addressing the following issues. (1) Stanford streamlined the set of annotation classes (i.e., the classes that implement the sharing of content between different annotators in the system. (2) The Baseline NL Processor can now run multiple instances of itself in parallel without duplicating shared annotators (e.g., one can run one processor with different parsers but the same named entity recognizer). (3) Stanford improved the strategy to combine multiple named entity recognizers in the same Baseline NL Processor instance. (4) Stanford integrated their statistical co-reference resolution system as a distinct annotator. (5) Stanford implemented a new deterministic co-reference resolution module.

Stanford added to the Baseline NL Processor an information-extraction system that extracts entity and relation mentions in the NFL domain. Overall, this achieves an F1 score of 82.0 for entities and 68.0 for relations.

4.3.5 Use Cases and Evaluation

The Phase 1 evaluation was on machine readability (predicting readability judgments; humans vs. machine parsing). Stanford took primary responsibility for the machine readability task on the FAUST team and completed a version of a readability assessment system for the Phase 1 evaluation. Stanford implemented a baseline supervised classifier for readability detection, using an n-gram model and a logistic regression classifier.

In initial work, Stanford implemented a ratings model that detects the five-point readability scores as defined by LDC. They evaluated this classifier using a metric that penalizes the model if it generates readability scores that are far from the LDC scores (i.e., *perExampleAccuracy* = 1 - abs(*trueRating* - *systemRating*)/4). According to this metric, the accuracy of the ratings classifier is 92%. This is a strong indication that the readability assessment task, as currently defined, is not realistic. Based on these observations, LDC proposed a new version of this task.

The baseline classifier was then extended with new parse-based features (such as parse tree depth, probability of the tree, etc.) and surface features (such as number of characters per word, number of words per sentence, etc.). However, this work was not sufficient to satisfy the requirements of the revised readability task.

Stanford performed extensive error analysis on the outputs of our initial readability system. As a result of this, Stanford supplemented the readability predictor with a number of new features including additional parse features (local subtrees and syntactic heads); new discourse features (using co-reference, named entity, and discourse connectives); lexicalized features (lexical cohesion and average inverse document frequency); character-based features (including word prefix and suffix with the intention of handling the MT/non-MT distinction); pairwise features (conjoining all possible pairs of features but only keeping the ones which correlated well with the training data); detecting specific text post-processing errors made in machine translation systems (e.g., mismatched quotations, doubled periods); modeling document/paragraph structure; formality of text on the Internet (uppercase text, etc.); character and discourse connective-based language models; and several new syntactic patterns.

In addition to linear regression and classification, Stanford explored several new models for predicting judgments: SVM regression, SVM bagging, K-Nearest Neighbor regression, and correlation-based regression. The latter is a linear regression model that uses a loss function designed to approximate correlation errors directly.

Ultimately, using various statistical tests, Stanford handpicked features from all possible features to build four distinct predictors (two SVM regressors and two linear regressors) and then combined them by averaging their outputs. The resulting system satisfied the performance requirements for a system for assessing machine readability.

Phase 1 required preparations for the NFL use case to be tested in Phase 2. Wisconsin conducted several experiments with this use case. The results of these preliminary evaluations suggested using domain knowledge in guiding ILP search. Specifically, domain knowledge pertaining to general knowledge about English (from the Stanford NLP Toolkit); knowledge about temporal statements (e.g., “this week” and “next month”) that Wisconsin encoded, as well as general knowledge about the NFL, such as team names and nicknames, also encoded by Wisconsin.

4.4 PHASE 2 RESULTS

Our work in Phase 2 focused on the following two tracks:

(1) We developed state-of-the-art NLP and applied it to the use cases that were selected by DARPA for the Phase 2 final evaluation: NFL and IC. We continually improved our evaluation system that was used on the Phase 2 final evaluation. The final product is depicted in Figure 5, which shows how information flowed through our evaluation system during the evaluation.

This track involved SRI coordinating the efforts of Stanford and secondarily, Wisconsin and SRI's PCE team on the NFL Use Cases; and of UMass and secondarily UIUC on the IC Use Cases. We worked with the Government Evaluation Team (ET) on a number of issues in making the FAUST results compliant with the ontology, DSRS, and MRAPI requirements (and improving those requirements).

- NFL: The Stanford module was evaluated, and we conducted an independent effort by the University of Wisconsin that was evaluated internally, but not in the MR program evaluation.
- IC: The University of Massachusetts module was evaluated and used Stanford NLP modules. We also conducted an independent effort by UIUC that was evaluated internally, but not in the MR program evaluation.
- As part of this track, our team explored various schemes of JI among NLP components.

(2) The second track was the continuing design and implementation of the FAUST reading system, a system that incorporates modules for performing probabilistic inference based on a small set of hand-engineered probabilistic rules and the output of various NLP modules. Research centered on explorations of more global JI schemes. One focused on Wisconsin's combination of MLN-style inference and inductive logic programming, and the second on SRI's MLN-based Probabilistic Consistency Engine. In addition, a variety of such schemes were developed and explored as described later in this section.

In the remainder of this section, we report our results organized by the major tasks. Multiple institutions contributed to each task, and some institutions contributed to several tasks.

4.4.1 Natural Language Processing

4.4.1.1 Stanford

Stanford implemented a joint NER (Named Entity Recognition) and LDA (Latent Dirichlet Allocation) model. State-of-the-art models for NER such as CRFs typically operate at the sentence level and use only lexical and morphological information for identifying the mentions of named entities in text. Therefore, these models miss the information on the topical context of the document that a given sentence is part of. Stanford's hypothesis was that the topical context of the document can sometimes help disambiguate the correct entity type of a given mention (e.g., the occurrence of the name "Washington" in a document that discusses politics could imply that the mention is a reference to Washington D.C., a LOCATION; while the same name mentioned in the context of movies could be referring to Denzel Washington, a PERSON).

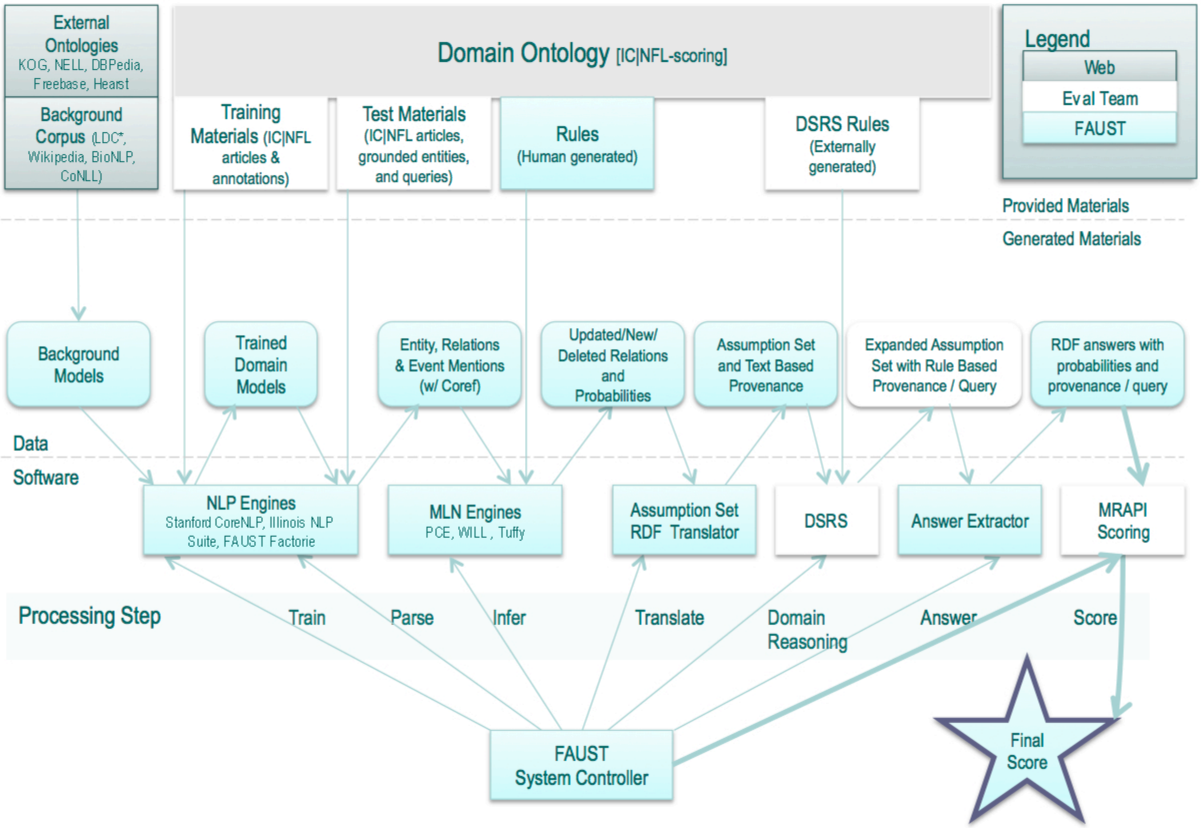

External Ontologies
KOG, NELL, DBPedia, Freebase, Hearst
Background Corpus (LDC*, Wikipedia, BioNLP, CoNLL)
Domain Ontology [IC|NFL-scoring]
Training Materials (IC|NFL articles & annotations)
Test Materials (IC|NFL articles, grounded entities, and queries)
Rules (Human generated)
DSRS Rules (Externally generated)
Legend
Web
Eval Team
FAUST
Provided Materials
Generated Materials
Background Models
Trained Domain Models
Entity, Relations & Event Mentions (w/ Coref)
Updated/New/ Deleted Relations and Probabilities
Assumption Set and Text Based Provenance
Expanded Assumption Set with Rule Based Provenance / Query
RDF answers with probabilities and provenance / query
Data
Software
NLP Engines
Stanford CoreNLP, Illinois NLP Suite, FAUST Factorie
MLN Engines
PCE, WILL, Tuffy
Assumption Set RDF Translator
DSRS
Answer Extractor
MRAPI Scoring
Processing Step
Train
Parse
Infer
Translate
Domain Reasoning
Answer
Score
FAUST System Controller
Final Score

Figure 5: FAUST Phase 2 system showing data flows as run for the Phase 2 Evaluation.

Stanford refined its work on deterministic "sieve-based" co-reference into a highly modular, deterministic, sieve-based co-reference solution system and released it as part of the Stanford CoreNLP distribution. Most co-reference resolution models determine if two mentions are co-referent using a single function over a set of constraints or features. This approach can lead to incorrect decisions as lower precision features often overwhelm the smaller number of high precision ones. To overcome this problem, Stanford created a simple unsupervised co-reference architecture based on a sieve that applies tiers of deterministic co-reference models one at a time from highest to lowest precision. Each tier builds on the previous tier's entity cluster output.

Further, the model propagates global information by sharing attributes (e.g., gender and number) across mentions in the same cluster. This cautious sieve guarantees that stronger features are given precedence over weaker ones and that each decision is made using all of the information available at the time. The framework is modular: new co-reference modules can be plugged in without any change to the other modules. In spite of its simplicity, the approach outperforms many state-of-the-art supervised and unsupervised models on several standard corpora. The latest sieves implemented by Stanford incorporate semantic information either from WordNet, Wikipedia, or Freebase. The first sieve links two mentions if their attributes agree and the WordNet path between their corresponding WordNet synsets is less than a threshold. The second sieve discovers name aliases using WordNet synsets, Freebase alias slots, and Wikipedia links. Stanford also implemented an additional method that uses discourse structure information of a document. By finding and using the speaker information of discourses, the system can do co-reference resolution better, especially for pronouns. Stanford has mostly used gold mention boundaries for this task, but recently Stanford has implemented the first version of rule-based mention detection to achieve more realistic end-to-end co-reference system.

Stanford continued work on their systems for slot-filling and entity-linking based on distant supervision for participation in the slot-filling and entity-linking tasks in the TAC-KBP shared task. The approach involves using distant-supervision from Wikipedia/DBpedia, Freebase, and snippets from Google search results.

Stanford made the following improvements to their slot-filling system [89]:

- Stanford has been developing error-analysis software for their TAC-KBP 2010 slot-filling system, primarily to better identify sources of recall errors.
- All of the corpora (TAC knowledgebase, Wikipedia, and web snippets) have been parsed and indexed. This should enable models to obtain higher recall scores. Considerable CPU time were spent preprocessing the 1.7 million documents in the KBP dataset with the full Stanford NLP pipeline (including parsing and co-reference) annotations to enable for faster systems which can utilize larger portions of the training data.
- Stanford explored whether domain-adaptation techniques (e.g., Hal Daume's Frustratingly Easy Domain Adaptation method and baselines) can be applied to inconsistencies in the three corpora.
- Stanford implemented a parallel version of their KBP slot-filling system that reduces training times by one order of magnitude. The speedup comes from distributing the search for examples in Stanford’s distantly supervised system.
- Stanford implemented several extensions of the KBP slot-filling system: (1) The system now supports both a mention model (where each mention is modeled as a separate datum) and a relation model (where all mentions of the same slot are merged into a single datum); (2) The system now supports both multiclass and one-vs.-all classification models; (3) The features used for slot extraction have been improved.
- Work was also performed on the automatic detection of trigger words (words which typically indicate a specific relation). While more complex measures such as pointwise-mutual information do not seem to work well for this task, simpler count-based measures have yielded good results.

Stanford made the following improvements to their entity-linking system [90]:

- For the entity-linking system, Stanford built an updated dictionary mapping strings of text to the potential Wikipedia pages that they can refer to. The dictionary includes frequency statistics and is the core of the Stanford entity-linking system. This dictionary can also be used as a resource independently of the entity-linking system.
- The entity-linking system has been refactored to better integrate with the Stanford CoreNLP pipeline and to enable easier future experimentation. For instance, the entity-linking system can now also take into account co-reference and NER features from the Stanford CoreNLP pipeline.
- The entity-linking system has also been improved with better error analysis and to support the use of different classifier types, such as SVM and logistic regression, multiclass and one-vs.-all classification. The entity system now supports several different strategies for linking a string to an entity (most frequent sense, heuristic, classifier trained using distant supervision, and context similarity matching).
- To help with the entity-linking effort, Stanford built an article classifier that classifies a Wikipedia page as one of a set of named-entity types (person, location, organization). The output of this can be used as a feature within entity linking. Stanford started to extend the system to work with more named-entity types, and created an initial dataset for evaluating the classification of fine-grained types.

Stanford developed an event-extraction system based on dependency parsing. To perform event extraction, Stanford framed it as a graph-learning problem. This enables a more natural representation of hierarchical events than previous approaches as well as the possibility of modeling inter-argument dependencies and naturally extending to recognizing events that span multiple sentences. The model now includes many of the features from state-of-the-art systems as well as additional features for capturing higher-order dependencies. Recent work has focused on converting the parser to an n-best parser with a reranker. A reranker enables the use of global features, which have previously not been captured by event extraction systems. Results are currently competitive with state-of-the-art systems given the minimal level of domain-specific feature engineering. Additionally, Stanford has been exploring different formulations for how to do document-level parsing but so far has not been able to obtain an improvement over sentence-level parsing.

- Efforts focused on an error analysis that indicated that a large source of errors came from insufficient robustness in the trigger detection system. Stanford was able to improve robustness by creating an ensemble model of several simpler trigger-detection systems. Stanford also has worked on improving the features both by including some more domain-specific features and by performing a new method of feature selection.
- Stanford continued extending the system to work on the BioNLP 2011 shared task [66]. This includes two new datasets and all domain-dependent code and domain-specific code has been factored out. Collaborating with UMass, Stanford and UMass produced a joint system as submissions to the shared task exploring several approaches for performing model combination [67]. The joint system placed first in three of the four tasks and second in the other task, outperforming the two individual submissions. The individual submissions also performed well.

Stanford worked to improve joint scenario-template learning and slot filling (as in the MUC task). Existing Stanford work learns scenario templates by learning related events and the semantic roles (slots) that characterize the template. In MR, Stanford extended the work to learn generalized MUC templates and to create a MUC information-extraction system. Stanford successfully built a system to perform the MUC task that outperforms older rule-based systems, and achieves results similar to previous weakly supervised systems. This work demonstrates how learning template structure from both a domain-specific and a broader corpus can complement each other.

Stanford has been using techniques from deep learning for large-scale joint-inference. Using recursive neural-network architecture for jointly parsing natural language and learning vector-space representations for variable-sized inputs, multiple tasks in natural language parsing (e.g., part of speech tagging, parsing, and paraphrasing) can be learned simultaneously. The core of the system includes context-sensitive recursive neural networks that can induce distributed feature representations for unseen phrases and provide syntactic information to accurately predict phrase-structure trees. The representation of each phrase can also be used as part of a paraphrasing system. For example, the phrases "decline to comment"' and "would not disclose the terms" are near each other in the induced embedding space. The current system achieves an unlabeled bracketing F-measure of 92.1% on the *Wall Street Journal* dataset for sentences up to length 15. Finally, this project uses GPUs to take advantage of their potential for massive parallelization. The same system has been used for paraphrase detection and improves a system by Chris Callison-Burch by 22% in F1 score. Further, the same learning architecture has been shown to perform well in other modalities such as image understanding. Stanford looked at more complex paraphrase-detection datasets as well as at sentiment detection. Stanford finished the derivations and implementations of a recursive autoencoder for sentiment prediction. The model achieves a new level of state-of-the-art performance on commonly used datasets and can better capture human sentiment in text than other models [95].

Stanford continued work on doing distant-learning for fine-grained named-entity extraction. The project has two different approaches, both of which can be used as input to downstream components. The first is flat and classifier-based, while the second is hierarchical and uses parsing. The former is easier and faster to train and for annotation, but it cannot model nested structure within named entities. To enable training on large datasets (e.g., Wikipedia), the conditional random field system was extended to use approximate inference (stochastic gradient descent). On the parsing approach, Stanford is planning to create a small dataset annotated with trees representing nested named entities. Determining a good representation of these is critical both for obtaining good inter-annotator agreement as well as for offering structures that can easily be learned by constituency parsers.

Stanford worked on extracting the focus, technique, and application of scientific papers from their abstracts. Currently, the patterns to extract these phrases are hand coded, but Stanford is working toward using machine-learning techniques to do so. One application of this information extraction is seeing the dynamics of sub-communities in a scientific community, whether they are focus-, technique- or application-centric over the years. When the patterns are automatically learned, they may be applicable to other domains as well.

4.4.1.2 UIUC and UMass

UIUC developed a textual entailment corpus with detailed per-phenomenon annotation. UIUC collaborated on further developing this corpus both with PARC and with an Italian team.

UIUC continued work on developing extensions to the existing semantic role-labeling package with the goal of extending SRL beyond verb predicates to a number of other relations. The focus is on integrating nominal relations, verb-based relations, and prepositional-based relations. UIUC investigated several global inference approaches to support this process.

UIUC continued work on improved named-entity resolution and co-reference resolution. UIUC developed a new ILP (integer linear programming) formulation for co-ref, as well as several new training algorithms to explore mention-based approaches, entity-based approaches, and joint learning approaches [63].

UIUC continued work on wikification: mention identification, disambiguation, and mapping of mentions to the appropriate page in Wikipedia. UIUC compared global approaches that simultaneously consider multiple mentions to local methods. The goal is using the wikification approach in resolving co-reference within and across documents. UIUC developed an approach to knowledge acquisition via the use of the Wikifier.

UIUC worked on a better mention-detection approach, to aid in co-reference, named entities, and wikification. The emphasis was on incorporating this module in other tools in a modular way, without affecting the trained model, to support a move to a new domain.

UIUC continued work on relation recognition and on event recognition, tracking, and de-duplication. For relation recognition, the current focus is on the IC domain. The current approach emphasizes an integration of the mention-detection process with the relation classification. This approach makes use of global inference over multiple components, including a Wikifier, co-reference resolution, and enforcing coherency constraints among relation types. UIUC is focusing on an approach that is minimally supervised and is beginning to incorporate indirect supervision into the training process.

A second effort in relation recognition focuses on a different type of analysis of relations. UIUC observed that a second dimension to the relation-extraction (RE) problem that is orthogonal to the relation type dimension exists. UIUC showed that most of these second dimensional structures are relatively constrained and not difficult to identify. UIUC proposed a novel algorithmic approach to RE that starts by first identifying these structures and then, within these, identifying the semantic type of the relation. In the real RE problem, where relation arguments need to be identified, exploiting these structures also enables reducing pipeline propagation errors. UIUC showed that this RE framework provides significant improvement in RE performance. A paper on this work appeared in ACL’11 [58].

UIUC started a new effort in event recognition, focusing on the identification of events and supporting event, and the identification of causality relations among them. In particular, UIUC developed some capabilities to identify discourse relations to support the work on events.

UIUC continued work on transfer learning and adaptation algorithms as a way to enable existing NLP tools to generalize better to data that is different from the training data.

UMass researchers developed improved methods for constituency parsing and its integration with named-entity recognition. By expressing a labeled bracketing model in a factor graph, accurate parsing can be achieved without consulting a large grammar or, building on UMass work in 2010 on marginal relaxation, by using only a small number of grammar-rule constraints.

UMass developed new techniques of unsupervised domain adaptation based on information-retrieval models for context aggregation and applied it to improved named-entity recognition.

UMass compared their unsupervised generative models for relation extraction against DIRT and USP, which are two similar unsupervised methods for clustering textual expressions of the same semantic relation. They demonstrated that in a distantly supervised relation-extraction framework, the unsupervised generative model features could improve performance more than DIRT and USP. Their results appeared in an EMNLP paper [35].

UMass worked on jointly resolving entities of different types (people, authors, papers, and venues) with minimal supervision over Bibtex-style Mongo-db records. An initial version of the framework was implemented and scaled to large quantities of Bibtex records. UMass investigated ways of incorporating human edits to this data in a probabilistic manner by treating them as evidence in a graphical model.

UMass worked on the generation of labels for a cross-document co-reference dataset by using Wikipedia entities to generate web search queries.

UMass summarized their work on biomedical event extraction in an EMNLP paper [76], which presents three models of increasing complexity. The first one matches state-of-the-art and requires only O(n^2) time. The final one uses dual decomposition for inference and outperforms the state-of-the-art substantially. The system was also extended to Task 2 of the shared task, leading to almost 9% F1 improvement over results from the BioNLP 2009 shared task.

UMass and Stanford wrote a FAUST paper for the 2011 BioNLP shared task [67]. The final submission was based on the dual decomposition model of UMass stacking the event-parsing model of Stanford. The UMass and Stanford FAUST team ranked first in three of the 2011 BioNLP shared tasks.

UMass continued to re-implement a version of their ECML and EMNLP work on relation extraction with distant supervision. This version is based on FACTORIE and will scale up to the whole the *New York Times* corpus. UMass incorporated co-reference annotations (within and across documents) into their system.

UMass worked on large-scale joint co-reference, segmentation, and alignment for semi-structured data. The goal is to cluster records and labeled/segmented texts without any supervision. UMass used a conditional random field model with posterior constraints for inference/learning. Significant progress was made in terms of speed of inference. On one dataset, a significant performance improvement was observed via joint co-reference and segmentation.

UMass continued to develop and implement a unified approach to both the bootstrapping of relations and distant supervision, based on the framework of Posterior Regularization. They found that in expectation, constraints such as “at-least-once” used in previous works are substantially harder to deal with. For example, relation with 1000 mentions of which not one has higher probability than 0.1 still is almost certainly active.

4.4.1.3 PARC

PARC finalized and delivered to SRI BD-1, custom data storage for annotated text.

PARC worked on identifying and processing the spatial information contained in texts for reasoning. PARC proposed an annotation of spatial paths to the ISO-space Workshop (Airlee Center, 2010). The meeting was organized by James Pustejovsky (Brandeis) and brought together an international group of logicians, database specialists, computational linguistics, ontologists, and NGA representatives to discuss the format of an annotation scheme for spatial information. PARC's proposal argues for a restructuring of the proposed ISO annotation for paths and extends PARC's own representation for paths. It gives a unified representation of static and dynamic locational and directional paths, bringing out the similarities between sentences such as "The road went from Palo Alto to Menlo Park" and "John went from Palo Alto to Menlo Park," as well as those between "The road went to the right/the west" and "John went to the right/the west." This enterprise is of obvious interest to MR, as information about what or who is where or what is happening at the particular place is essential to understand text.

PARC has a sophisticated rule-based system to calculate intervals that help locate events in time and infer temporal relations. To enable the integration of the PARC system into the Stanford pipeline (or any other pipeline used within FAUST), PARC worked on a NER-like recognizer that identifies temporal modifiers and tags the corresponding strings with the structure necessary to compute the relevant semantic representation. This enables inferences about interval relations even when there is no calendrical anchoring. The annotation categories are based on the interval calculations done in the PARC rule-system but the idea is implemented as a cascade of FS annotations that can be run separately on text or be integrated with various parsers.

PARC, in previous work, provided a classification of lexical items and multiword constructions that give rise to two types of textual inferences: presuppositions (factives) and entailments (implicatives). During this period, work extended in two directions:

(1) PARC looked at a particularly interesting case of phrasal implicatives involving verbs of reckless spending such as "waste," "blow," and "squander." These constructions are two-way implicatives. What makes this class interesting is that the polarity of the entailment changes depending on the type of noun in the construction. For example,

waste time to X => X   not waste time to X => not X
waste chance to X = not X   not waste chance to X => X

In this construction, "occasion" nouns such as "chance" and "opportunity" contrast with "resource" nouns such as "time," "morning," "energy," and "ability." PARC constructed a list of nouns that frequently occur in construction with verbs of reckless spending. The semantic class of resource nouns in particular is not to be found in WordNet or other available lexical resources or ontologies.

(2) PARC investigated constructions that are a source of systematic pragmatic inferences. Pragmatic inferences are different from presuppositions and entailments in two ways. They may not be contrary to facts already known, and they can be explicitly denied by the speaker without a contradiction. A case in point is the construction "X meant to Y" and its negation "X did not mean to Y." When looking at a large collection of naturally occurring examples, clearly most of them are consistent with this inference pattern:

X meant to Y   ==> X did not Y
X did not mean to Y ==> X did Y

However, these inferences are not based on logical entailment. There is no contradiction in “The boy really meant to call his friend. It was no accident.” In the negative case, the inference may be cancelled explicitly with no contradiction “The boy did not mean to wreck the car and he managed to prevent the accident.” The “X mean to Y” case is an example of a large class of constructions that give rise to pragmatic inferences. While much anecdotal discussion of these cases exists in the linguistic literature, PARC is doing the first systematic survey of these constructions.

PARC worked on relating their implementation of a monotonicity calculus, polarity propagation, and inferring semantic relations to Natural Logic. The aim was to extend the notion of Natural Logic as a subcomponent of MR reasoning by precomputing some generalized entailment relations. For example, after inferring that peaceful protest entails protest, the system should further infer that every protest entails every peaceful protest. Two of the outstanding issues are (1) how local semantic relations are and (2) what type of representation the generalized entailment calculation should be done on. Temporal modifiers and the interaction of presuppositional expressions, such as factive predicates, with negation show that the calculation of generalized entailment should be done on a more normalized representation that is derived from the syntactic parse, and that semantic relations cannot be strictly local.

PARC explored Natural Logic as a framework to pinpoint the formal elements that enable inference directly from textual analysis, and to delimit precisely the lexical classes that allow an inference to go through. As part of this effort, PARC started to organize a Workshop on Inference from Text that was held at the LSA-Institute in Boulder in 2011. The aim was to bring together people involved in MR and other groups working on NLU with semanticists and logicians to foster collaboration, with a special focus on developing proof-theoretical approaches. PARC organized another workshop at CSLI (Stanford) focused on proof-theoretic, and especially natural logic, approaches to (computational) semantics.

PARC worked on understanding the inferences associated with propositional attitude predicates that describe acts of communication intended to influence the future actions of the agents to whom the act is directed. The relevant predicates, which occur in the IC corpus, include those describing commands, suggestions, requests, warnings, etc. A common core to these inferences is the notion of an agent’s commitment to a belief or a preference for action. PARC’s aim is to use the results of these theoretical studies as the underpinnings of a lexicon of these predicates marked with their inferential properties.

PARC looked at how inferential information about a domain can be constructed and used. In collaboration with SRI and the University of Wisconsin, PARC explored how the inferences in the DSRS might be used by their JI engines. In addition, PARC looked at how inference rules expressed in NL can be transformed into a logical form to be used by an inference engine. The use of English statements of inference rules would leverage learning from reading to facilitate transition to new MR domains.

PARC extracted information from RDF schemas into their system to demonstrate how such information can be exploited in the interpretation and disambiguation process. This approach can be used by other systems used within FAUST and enables exploiting RDF to make domain information available.

4.4.2 Joint Inference

4.4.2.1 SRI and Wisconsin

SRI continued to work, with Wisconsin, on devising and implementing the Anytime Lifted Belief Propagation (ALBP) algorithm. This version is the most general LBP algorithm presented to date, dealing with Context-Sensitive Independence (CSI), symbolic unknown quantities, and compact constraint representations. In addition, they assisted in the formalization of rewriting rules in Higher Order Logic, which led to a rigorous framework for proving its correctness and studying theoretical properties of the algorithms. For example, we can show that the symbolic evaluation implementation of LBP leads to an algorithm identical to the original LBP by Domingos, but with better representation of intensionally defined sets. In addition, it can be easily generalized to handle CSI and interpreted functions.

The SRI Probabilistic Consistency Engine (PCE) team created MLNs in for the evaluation domains, performed experiments to test whether PCE inference is able to improve the results given by the FAUST NLP engine. This effort resulted in several interesting research problems related to the theory/modeling of PCE (or in general Markov Logic Networks) that are currently being explored by the PCE team. They extended their end-to-end proof-of-concept demo, automating the whole flow (from NLP extraction using Stanford's BaselineNLProcessor to final inference using PCE, in the NFL domain) to support the Common Annotation Format (CAF). They implemented a weight-learning module for PCE and designed an algorithm for learning confidence scores of data sources in PCE and another algorithm for combining evidence from multiple sources to support JI in FAUST. This work was described in a ECML PKDD 2011 paper [88].

Wisconsin developed RDN-BOOST, a functional gradient boosting algorithm that can efficiently learn both the structure and parameters of Relational Dependency Networks (RDNs). RDNs are graphical models that extend dependency networks to relational domains where the joint probability distribution over the variables is approximated as a product of conditional distributions. Unlike other learning approaches for RDNs, which learn a single probability tree per random variable, RDN-Boost learns a series of relational function-approximation problems using gradient-based boosting. In doing so, Wisconsin can easily induce highly complex features over several iterations and, in turn, quickly estimate quickly a very expressive model. This approach has several novel features:

- Structure and parameters of the model are learned simultaneously
- A Gibbs Sampling algorithm that can perform JI over several queries simultaneously
- Learning can be performed in cases of hidden or partially missing data

RDN-BOOST can be found at http://pages.cs.wisc.edu/~tushar/rdnboost/. A paper describing the technical details of this work appeared in the Machine Learning Journal [98][5].

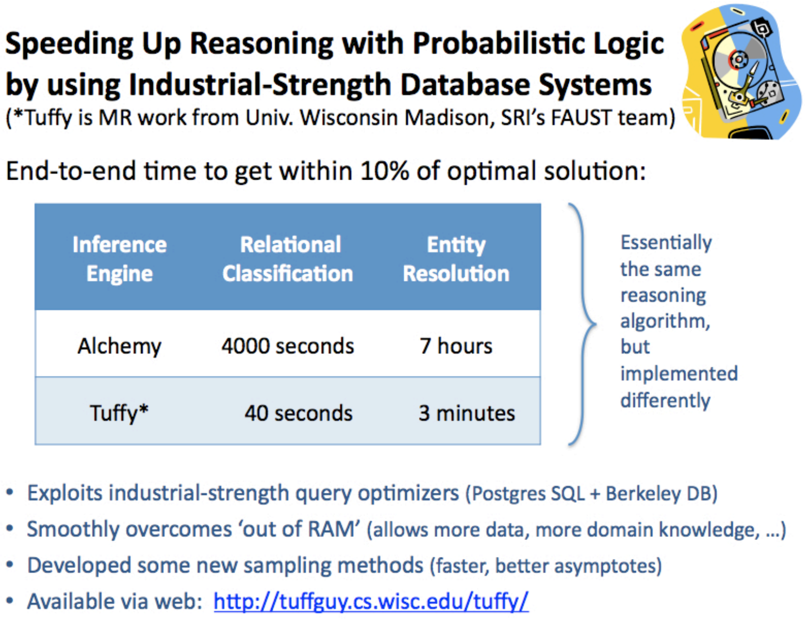
[5] available at http://ftp.cs.wisc.edu/machine-learning/shavlik-group/natarajan.mlj12.pdf

Wisconsin made several improvements to Tuffy, their probabilistic, deductive database-management system (http://www.cs.wisc.edu/hazy/Tuffy/). Significant changes include:

- *Marginal inference*: Tuffy is now capable of estimating the marginal probabilities of queries (marginal inference) in addition to finding the most likely assignment of all the random variables. Tuffy implements the MC-SAT algorithm to perform marginal inference.
- *Weight learning*: Tuffy is also capable of learning the weights on the various Markov Logic rules. This is done via the implementation of pre-conditioner-scaled conjugate gradient descent to perform discriminative weight learning using samples from MC-SAT.
- *Datalog rules and functions*: Tuffy can execute Datalog (declarative logic) rules and functions, which are useful features for writing MLN programs for Machine Reading tasks.

The results on Tuffy are summarized in Figure 6, and its architecture and technical details were described in a VLDB 2011 paper [87][6].

Wisconsin initiated development of FELIX (previously named Mobius, but renamed due to a name conflict) that extends Tuffy to both enhance scalability and increase the quality of performance on machine-reading tasks on the NFL test bed. While sophisticated statistical-reasoning frameworks (e.g., MLNs) have demonstrated impressive quality on small information extraction (IE) tasks, they currently do not scale to enterprise-sized tasks. The key bottleneck of their scalability is their monolithic approach to inference: not only do these frameworks express all IE subtasks in one program, but they also try to solve all subtasks using a single algorithm. Many of these subtasks (e.g., classification, co-reference resolution) have specialized algorithms (e.g., Viterbi, classical graph algorithms) with both high performance and high quality. To address this, Wisconsin is developing an operator-based system, FELIX, that takes the same sophisticated IE program (e.g., an MLN), decomposes it into multiple parts, and then solves each part with specialized statistical operators [85].



Speeding Up Reasoning with Probabilistic Logic by using Industrial-Strength Database Systems
(*Tuffy is MR work from Univ. Wisconsin Madison, SRI's FAUST team)

End-to-end time to get within 10% of optimal solution:

Inference Engine	Relational Classification	Entity Resolution
Alchemy	4000 seconds	7 hours
Tuffy*	40 seconds	3 minutes

Essentially the same reasoning algorithm, but implemented differently

- Exploits industrial-strength query optimizers (Postgres SQL + Berkeley DB)
- Smoothly overcomes 'out of RAM' (allows more data, more domain knowledge, …)
- Developed some new sampling methods (faster, better asymptotes)
- Available via web: http://tuffguy.cs.wisc.edu/tuffy/

Figure 6: Wisconsin Phase 2 results for joint inference.

[6] Available at http://hazy.cs.wisc.edu/hazy/papers/tuffy-vldb11.pdf

4.4.2.2 Other Team Members

The University of Washington worked toward a new end-to-end solution to machine reading that builds on top of USP and enables efficient large-scale joint inference. In Phase 2, UW completed work in three synergistic directions toward this goal:

- First, UW worked on a sounder, simpler, more scalable, online version of USP. UW made a step forward by developing a new deep architecture called the sum-product networks (SPNs) [72,84]. SPNs are more general than arithmetic circuits and enable efficient exact inference (linear in the network size). UW successfully developed the basic algorithms for SPNs and began building an online version of USP based on SPN, which was called USPN. USPN uses a SPN to process text sequentially as a character stream, with the meaning of the read portion compactly represented as a deep network. UW started to develop the theory of USPN.
- Second, UW worked on unifying probabilistic and logical inference via a new approach called probabilistic theorem proving (PTP) [73], as well as "Approximation by Quantization" (ABQ), a new approach for efficiently conducting approximate probabilistic inference [74]. PTP is a unified approach to JI that enables a seamless integration of lifted inference and Monte Carlo methods. PTP splits on full first-order logic formulas at each step, greatly reducing the time and memory required for inference, particularly when long formulas or long inference chains are involved. UW completed exact PTP and developed an approximate version based on importance sampling. The goal of ABQ is to scale up JI by exploiting both approximate, context-specific independence and approximate determinism.
- Third, UW investigated inference methods that could benefit from ontological lifting and how one may apply Ontological Lifted Probabilistic Inference (OLPI) to those algorithms. In Phase 1, UW created OLPI as a general framework for ontological lifting that can be applied to many inference algorithms. UW began developing a general coarse-to-fine probabilistic inference framework based on arithmetic circuits. Arithmetic circuits provide an efficient framework for exact inference given a model with many context-specific independencies, such as models containing knowledge extracted from text. However, large models may still compile into circuits that are too large for the available time and memory. UW's inference algorithm provides a structured way to approximate inference with arithmetic circuits given realistic time and memory constraints. UW also developed a new coarse-to-fine paradigm based on hierarchical mean-field approximations.

Columbia University continued its work on JI using dual decomposition and Lagrangian relaxation. One focus was on an algorithm for non-projective dependency parsing. Simple models for non-projective dependency parsing, "arc-factored" models, can be decoded using directed spanning tree algorithms. The non-projective dependency-parsing problem is however known to be NP hard for practically any generalization to more complex models. Columbia developed a decoding algorithm for head-automata models for non-projective dependency parsing. Head-automata models enable a rich set of features in parsing. The method makes use of dynamic programming algorithms for decoding the set of modifiers for a particular head word, in combination with a directed spanning tree algorithm that enforces the constraint that a well-formed parse tree has to be a well-formed directed tree. The method produces exact solutions, with certificates of optimality, on the vast majority of test examples.

UIUC continued working on JI and Joint Learning for NLP and made exciting progress on new learning algorithms for learning structure with indirect supervision. Work continued in the direction of indirect supervision and learning latent structure in the context of relation extraction. Recent work developed a theoretical understanding and very significant gains in the efficiency of joint training.

UIUC developed a new, principled framework for performing efficient learning with declarative constraints over structured output spaces. This framework establishes a spectrum of learning algorithms ranging from global learning to independent models, enabling a customization of the hardness (and efficiency) of learning based on a given set of constraints. UIUC gives both theoretical results and show experimentally that these algorithms are robust and, while being significantly more efficient, provide performance close to global learning.

UIUC began developing, together with UMass, an abstraction of a JI module, that will allow us to integrate individually learned modules and run multiple forms of JI for it. The first test case for this would be the integration of the UIUC and UMass RE modules.

UMass further improved a proposal for a FAUST Joint Inference architecture based on decomposition and message passing. Here, messages can either be marginals (as in BP) or be MAP states (as in Dual Decomposition). FAUST is proposed to be divided into roughly four layers: low-level NLP; entity and relation-mention extraction; co-reference; and KB reasoning.

UMass integrated ideas from particle filtering into Metropolis-Hastings (MH)-based inference for graphical models. This method of inference was applied to query evaluation in large-scale probabilistic databases based on factor graphs. On entity resolution, the particle-based MH approach achieved a lower squared error than a baseline parallel chain MH method.

UMass continued work on probabilistic databases backed by factor graphs and MCMC inference and successfully demonstrated the ability to capture the type of high tree-width graphical models required for solving grand unified data integration problems. UMass began building probabilistic databases of ontology alignment, co-reference resolution, and record extraction in a scalable infrastructure that taps indexed key-valued stores in lieu of relational databases based on SQL.

UMass continued work with SRI's Hung Bui on lifting the marginal polytope. They evaluated lifted max product versus lifted MPLP (derived through the lifted marginal polytope) on two synthetic datasets. In both cases, lifted MPLP outperforms MPLP, MP, and lifted BP.

4.4.3 Learning for and from Reading

The University of Washington developed a new probabilistic graphical model for knowledge-based weak supervision (aka distant supervision) of relation extraction. The new model is especially appropriate for learning when the knowledgebase is only loosely aligned with the textual corpus (e.g., when only a small number of the mention-pair matches in the text actually correspond to true discourse about the relation at hand). UW's new model also enables overlapping relations, in contrast to previous models that assume that R1(a, b) precludes another relation, R2, also holding between a and b. UW tested their approach by learning a wide range of extractors over the *New York Times* corpus using Freebase as the KB for weak supervision. Their system is about 100x faster than previous approaches, has slightly higher precision, and substantially higher recall. A paper on this research was published at ACL 2011 [56], with many institutions downloading the source code and adopting the method in their work.

UW designed an extension of their KB weak supervision model to do named-entity recognition for a much larger set of entity types than are currently employed. They started work to combine this NER system with their relation-extraction system in a manner that enables joint entity and relational extraction.

UW built an ontology matcher, which searches the space of datalog expressions to find view descriptions of an ontology's relations in terms of the others. Preliminary experiments showed that the autonomous system can successfully match NELL relations to Freebase with high accuracy and showed some success matching IC-domain relations to Freebase.

UW built an ontological smoothing system that enhances a minimally supervised extractor learner by augmenting its training data using knowledge-based weak supervision from an autonomously matched relation (or join between relations) in Freebase. Preliminary experiments on NELL relations showed that reading performance could be improved substantially.

UW built a rule-learning system using a novel form of inductive logic programming optimized for noisy source tuples with minimal negative examples. They showed that the learned rules could dramatically improve the recall of open extraction with minimal loss in precision. A paper on this research was published at EMNLP 2010 [34].

Wisconsin continued implementing an approach for using Inductive Logic Programming (ILP) plus MLNs to learn patterns from the sample extractions provided for a domain. Wisconsin focused on using domain-specific background knowledge for learning the rules. In addition, Wisconsin provided information about the background knowledge and created algorithms for generating useful negative examples.

Wisconsin improved their Wisconsin Inductive Logic Learner (WILL) system that learns first-order logic rules for the NFL queries. This improved system used the standard formats for input and output and served as a baseline for rule learning against which all the other systems could be compared. In addition, Wisconsin improved domain-specific background knowledge regarding NFL games and teams, and negative-example generation approaches using entity mentions.

UIUC continued work on Transfer Learning and Adaptation algorithms as a way to enable existing NLP tools to generalize better to data different than the training data.

Specifically, UIUC developed a new approach to adaptation that makes use of both (1) abstraction of features and (2) adaptation to new definitions in the new domain. UIUC built on a new theoretical analysis of existing adaptation schemes and a result that showed the necessity of using both aspects in order to adapt.

UIUC developed new learning algorithms for Learning Hidden (latent) Structure done within the Constrained Conditional Models (CCM) framework. UIUC evaluated it on textual entailment, paraphrasing, and translation with very good results. The key application being studied within this framework is that of Generalization across Relations. UIUC developed a new learning algorithm for structured prediction that can be trained using an indirect supervision signal—a signal generated in a cheap way from a companion binary decision problem associated with the structure prediction problem.

UIUC developed a Taxonomic Relations classifier that makes use of stationary ontologies and Wikipedia to infer robustly is-a relations and sibling relations [99]. The approach makes use of global inference over multiple related relations.

UIUC worked on supporting integration of multiple levels of natural-language analysis. This includes a new multiview-based alignment algorithm, which incorporates multiple levels of analysis of natural language—including POS; shallow parsing; dependency parsing; semantic role labeling; named entities and co-reference resolution—as a way to align text and hypothesis in the context of textual entailment.

UMass had a technical breakthrough that dramatically reduced the asymptotic time complexity of Generalized Expectation training for linear chain CRFs. They exploited this by incorporating external knowledge from the web (i.e., from DBLP or Wikipedia) into learning with generalized expectation.

UMass worked on new methods for estimating probability of correctness of model predictions. This method has applications in accuracy estimation and semi-supervised and active learning.

UMass continued work on new methods for minimal effort evaluation of lightly supervised learning. These methods involve carefully selecting examples for manual evaluation.

UMass began work on applying constraint-based, semi-supervised learning to dense factor graphs, where exact inference is not possible. Preliminary experiments showed that lowering the temperature and sampling provided a good approximation for small graphs; however, the approach requires further investigation for larger graphs.

4.4.4 Infrastructure, Software Engineering, and Integration

SRI's SE Team built evaluation software and the necessary tooling to accomplish the Phase 2 evaluation tasks for both NFL and IC. The integration team leveraged the common annotation format (CAF) to provide common utilities to map between the representation of data for FAUST and the representation of data as defined by the Government evaluation team (ET).

The SE team integrated and tested the modules as released by our major subcontractors (Stanford, UMass, UIUC, and Wisconsin), as well as the Probabilistic Consistency Engine (PCE) from SRI's research team, which was used to augment answers in the NFL domain.

The SE team reviewed every answer from the Dry-Run Gold Standard and determined that the data was too full of errors to be generally useful.

The SE team created visualizations and other analysis tools to assist in the hand scoring of FAUST in preparation of retaking the evaluation.

Wisconsin developed and implemented *RDN-BOOST*, which extends dependency networks into relational domains and can represent a joint probability distribution over the variables as a product of conditional distributions. RDN-Boost can learn a series of relational function approximations through functional gradient boosting, and this ensemble has higher expressivity.

Wisconsin extended Tuffy to perform weight learning and marginal inference for MLNs. This enables Tuffy to infer marginal probabilities (marginal inference) of the queries themselves, as well as to find the most likely assignment of all the non-evidence queries (MAP inference). In addition, Tuffy is also able to learn the weights on MLN rules. Wisconsin integrated Tuffy into the Phase 2 release of FAUST.

Wisconsin provided SRI with a list of relations extracted over the Dry-Run NFL corpus for testing integration. These relations were extracted using RDN-Boost and Tuffy using the tools provided by Stanford.

Stanford continued work on relation extraction over the NFL domain. Stanford updated the NLP components with several additional NFL mention types and relations and worked toward improved models and light inference in their NFL annotator. The system was integrated at SRI to create an initial end-to-end system that interfaced with the domain specific reasoning system (DSRS) and Machine Reading application programming interface (MRAPI) Scoring services.

Stanford continued work on their supervised system to extract entities, relations, and events. This system was used in several projects including the BioNLP and TAC-KBP shared tasks. The system received some improvements including a mechanism to perform basic logical inference; better syntactic head finding; improved conversions to Stanford Dependencies; and various other changes for the NFL information extraction system.

Stanford publicly released the Stanford CoreNLP (formerly Baseline NL Processor), an integrated suite of natural language processing components including tokenization, POS tagging; named-entity recognition; parsing; and co-reference.

UIUC provided an improved version of the Curator [103] to SRI with improved versions of many of its existing NLP tools.

UMass maintained and improved the FACTORIE platform through bug fixes and performance improvements. They worked on a web frontend for FACTORIE based on the lift framework.

UMass provided updated versions of their FAUST-FACTORIE module for FAUST. They provided substantial improvements to the co-reference engine based on using entity types as input features. They optimized the use of negative data in ACE to achieve better recall for relation extraction.

UMass further enhanced FACTORIE as they implemented an initial version of belief propagation that uses values without referring to variables and enables for parallelization. The code was based on a generic message-passing interface.

UMass started working on a distributed system for belief propagation over multiple machines. They began implementing parallel residual-splash belief propagation.

4.4.5 Use Cases and Evaluation

The Stanford supervised relation-extraction system was used in January 2011 to perform and pass the two NFL use cases in the Phase 2 evaluation.

Wisconsin applied their RDN-Boost approach on the Evaluation NFL corpora. They used handwritten MLN rules with Tuffy (an implementation of MLNs using RDBMS technology) to find consistent game descriptions in every article. A set of predicted games is consistent if no team plays more than one game per week, and in all games the winner's final score is larger than the loser's. Wisconsin provided these consistent extracted relations to SRI.

Numerous updates were made to the UMass FAUST-FACTORIE module, which was used in the Phase 2 evaluation for the four IC use cases.

UIUC continued to improve their relation-extraction engine for the IC use case. This component was not ready for the Phase 2 formal test but will be in the Phase 3 FAUST system.

Using the lessons learned during the Dry Run evaluation, SRI's SE team was able to modify the evaluation version of FAUST to increase its robustness and improve interaction with the ET's provided DSRS service. The SRI SE team invested a considerable amount of effort working with the Government ET to debug their DSRS service.

The initial Phase 2 evaluation suffered from a few bugs introduced just before evaluation, and we were granted the opportunity to re-take the evaluation. SRI spent significant effort improving robustness as well as developing metrics to automatically assess the quality of each new release from a subcontractor. We had to spend much more effort than anticipated with the ET's software and automated scoring system, but this still resulted in poor quality dry run results. Finally, we abandoned the ET's tools and data and developed our own in-house systems and hand-annotated answer sets. The efforts by the SE team with FAUST subcontractors resulted in a passing score on the test. Complete results can be found in the reports of the Government evaluation team.

4.5 PHASE 3 RESULTS

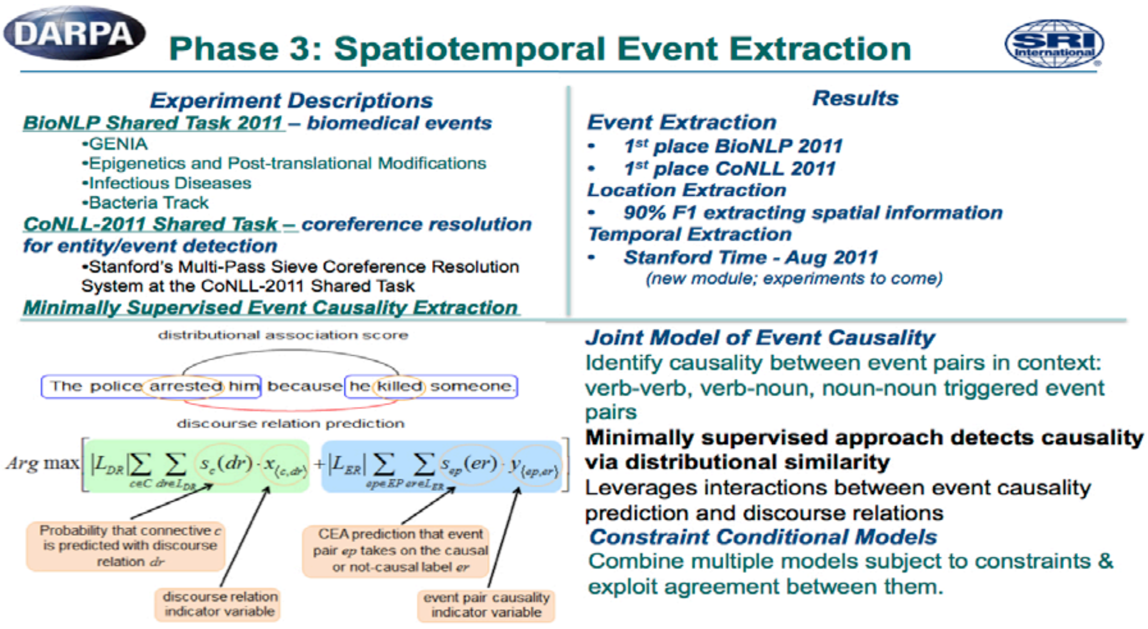
Our entire team participated in developing a detailed Phase 3 Research Plan, which involved collaboratively planning our approach to achieving Phase 3 goals. Our major Phase 3 activities were on the following tracks:

(1) Continuing to improve our evaluation system, which was used on the Phase 2 evaluation. The lessons learned were utilized to refine our efforts in Phase 3 for the spatio-temporal use case. This track involves both the MR-KBP and IC++ reading tasks for Phase 3, which are extensions to both the NIST sponsored TAC-KBP and the Phase 2 IC core reading task. Early in Phase 3, we successfully took and passed the Phase 2 retest in IC and conducted investigations in the Event Extraction Experiment (EEE).

(2) Continuing design and implementation of the FAUST Reading system, a system that incorporates modules for performing Probabilistic Inference based on a small set of hand-engineered probabilistic rules and the output of various NLP modules. This work was itself pursued on two tracks, one focused on Wisconsin's combination of MLN-style inference and Inductive Logic Programming, and the second on SRI's MLN-based Probabilistic Consistency Engine.

(3) Investigating a joint architecture developed by UMass and Illinois.

We report our results organized by the major tasks. Multiple institutions contributed to each task, and some institutions contributed to several tasks. In this section, we select three of our many results and present a figure describing each of them in more detail. These figures concisely summarize three of our most important results and were delivered to DARPA in 2011 as quad charts.


DARPA
SRI International

Phase 3: Spatiotemporal Event Extraction

Experiment Descriptions

BioNLP Shared Task 2011 – biomedical events

- GENIA
- Epigenetics and Post-translational Modifications
- Infectious Diseases
- Bacteria Track

CoNLL-2011 Shared Task – coreference resolution for entity/event detection

- Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task

Minimally Supervised Event Causality Extraction

distributional association score
The police arrested him because he killed someone.
discourse relation prediction
Probability that connective c is predicted with discourse relation dr
discourse relation indicator variable
CEA prediction that event pair ep takes on the causal or not-causal label er
event pair causality indicator variable

$$Arg\max\left[|L_{DR}|\sum_{c\in C}\sum_{dr\in L_{DR}} s_c(dr)\cdot x_{\langle c,dr\rangle} + |L_{ER}|\sum_{ep\in EP}\sum_{er\in L_{ER}} s_{ep}(er)\cdot y_{\langle ep,er\rangle}\right]$$

Results

Event Extraction

- ***1st place BioNLP 2011***
- ***1st place CoNLL 2011***

Location Extraction

- ***90% F1 extracting spatial information***

Temporal Extraction

- ***Stanford Time - Aug 2011***
  (new module; experiments to come)

Joint Model of Event Causality

Identify causality between event pairs in context: verb-verb, verb-noun, noun-noun triggered event pairs

Minimally supervised approach detects causality via distributional similarity

Leverages interactions between event causality prediction and discourse relations

Constraint Conditional Models

Combine multiple models subject to constraints & exploit agreement between them.

Figure 7: FAUST Phase 3 results for event extraction

4.5.1 Natural Language Processing

The results obtained by team members UMass and Stanford for event extraction are shown in Figure 7 [65,66,67] and are described in an ACL workshop paper [67][7]. Team members UMass and Stanford collaborated, and their module was first in BioNLP 2011 and CoNLL 2011 event extraction.

4.5.1.1 Stanford

Stanford continued work on their supervised system to extract entities, relations, and events, which is now used in several projects including the BioNLP and TAC-KBP shared tasks. They made improvements, including a mechanism to perform basic logical inference; better syntactic head-finding; improved conversions to Stanford Dependencies; and other changes for the NFL IE system. Stanford implemented extensions of the NFL IE system: (1) the generic NER is now better integrated with the NFL entities (for example, “second quarter” is a DATE in an open-domain environment, but not in the NFL domain); (2) the syntactic head detection heuristics for entity mentions (which are crucial for feature extraction) have been adapted to the NFL domain; (3) Stanford added a deterministic inference component that is capable of generating new NFL relations based on existing evidence (for example if Team A has a score of 10 and Team B has a score of 4, then Team A is the winner and Team B is the loser in the corresponding game). All three extensions substantially improved entity and relation extraction for NFL.

Stanford improved its sieve-based co-reference system, achieving state-of-the art results. Most co-reference resolution models determine if two mentions are co-referent using a single function over a set of constraints or features. This approach can lead to incorrect decisions as lower precision features often overwhelm the smaller number of high precision ones. To overcome this problem, Stanford created a simple, unsupervised co-reference architecture based on sieves that apply tiers of deterministic co-reference models one at a time from highest to lowest precision. Each tier builds on the previous tier’s entity cluster output. Additionally, Stanford’s model propagates global information by sharing attributes (e.g., gender and number) across mentions in the same cluster. This cautious sieve guarantees that stronger features are given precedence over weaker ones and that each decision is made using all of the information available at the time. The framework is modular: new co-reference tiers can be plugged in without any change to the other tiers. The approach outperforms many state-of-the-art supervised and unsupervised models on several standard corpora.

Stanford started work on a cross-document event-and-entity co-reference system [126]. Stanford implemented a document clustering system for preprocessing inputs for the cross-document co-reference resolution system, and continued to annotate a corpus for event and entity co-reference. In addition, Stanford finished annotating a corpus for event and entity co-reference.

Stanford used techniques from deep learning for large-scale JI. Stanford extended the recursive autoencoder and developed a new pooling technique for deep learning. The recursive autoencoder algorithms obtain state of the art performance on standard sentiment analysis datasets as well as the *Microsoft Research Paraphrase Corpus*. Stanford continued to develop new approaches for compositional semantics using recursive deep learning and explored learning multiple vectors for words in word vector space models. This model now obtains state-of-the-art

[7] http://stanford.edu/~mcclosky/papers/riedel-bionlp-2011.pdf

performance on human similarity judgments. Stanford developed a new technique for holistic compositionality that bridges the gap between formal semantics approaches and vector space models. A prototype for pre-training recursive models in an unsupervised way looks very promising, and Stanford has now scaled it up to large datasets. Stanford also showed that the new model can learn complex adverb-adjective relationships and got preliminary results for classifying noun-noun relationships.

To add temporal information to the Stanford CoreNLP pipeline, Stanford created SUTime, a Java library that recognizes and normalizes temporal expressions using deterministic patterns [101]. SUTime is similar in functionality to the Perl GUTime library. Stanford made a servlet that shows the results of both SUTime and GUTime for comparison.

Stanford worked on a probabilistic system for identifying and grounding time expressions into a representation compatible with the temporal slot-filling task. This project aims to expand on the range of expressions that are handled by GUTime, more elegantly handle ambiguity in the lexicon (e.g., “last week” vs. “last week of May”) and allow for training from an arbitrary time-expression tagged corpus. This grounding would be learned in a distantly supervised setting, trained on <phrase, grounded time> pairs while inferring the latent compositional structure. Unlike GUTime and SUTime, this system will be able to provide distributions over possible groundings. Stanford has built a preliminary system and is continuing to improve the parsing framework. Stanford also prepared for data collection of temporal expressions on the Amazon Mechanical Turk, focusing on expressions relating to meeting scheduling. Stanford evaluated the model on the TempEval-2 task against state-of-the-art systems, achieving comparable results.

Stanford worked on models to reconstructing timelines of significant events for interrelated entities using recent techniques from JI [125]. The approach is to encourage agreement between two distantly supervised models. The first model performs temporal extraction to map events to their time spans. The second models consistency between events and allows the learning of constraints and tendencies. Examples of these include “people typically go to school when they're 6–21 years old,” “people typically get married after attending at least one school,” “children are born when their parents are 20–40 years old,” and “people can't work at an organization until it has been founded.” By learning these patterns, the resulting timelines should be more consistent. JI combining these two models can be performed with Gibbs sampling. Ultimately, these models can be evaluated in the Temporal KBP or MR-KBP-style tasks.

- Stanford scraped 400,000 articles from Freebase as a source of distantly supervised temporal information. Stanford started to integrate their KBP module with the temporal-extraction module and created a visualizer to examine the results. The temporal information from Freebase has been aligned with the KBP knowledgebase. Using this information, Stanford created a baseline classifier to learn the mappings between events and their associated time spans. Candidate time spans are given labels such as “starts the event,” “ends the event,” and “no relation to the event.” Stanford started building the consistency model that will estimate the likelihood of two events co-occurring.
- The representations for the two models have been unified and a Gibbs sampler that incorporates information from both models was built. Some new features were added to the mapping and consistency models. Recently, Stanford has been exploring improving the performance of the joint model relative to its pipelined and heuristic baselines.

Stanford continued development of their slot-filling system for the TAC-KBP task based on distant supervision. Stanford rewrote and reengineered several part of its TAC-KBP slot filling system. Due to this reimplementation, the Stanford system now performs 30% (relative) than its previous version.

- The code for slot filling has been optimized to handle larger datasets. All of the corpora (TAC knowledgebase, Wikipedia, and web snippets) have been parsed and indexed. This has enabled models to obtain higher recall scores. The Stanford slot-filling system is a distributed system that can search for examples in parallel during training. The system also supports the following: (1) both a mention model (where each mention is modeled as a separate datum) and a relation model (where all mentions of the same slot are merged into a single datum); (2) both multiclass and one-vs.-all classification models.
- Stanford implemented several extensions to the slot-filling system: (1) the system can now extract slots from passages containing multiple sentences, instead of just one sentence; (2) the system can now run in bootstrapping mode, where successive iterations can clean up incorrect annotations in the “silver” data produced using distant supervision. Stanford participated in the TAC-KBP evaluation. Several extensions of the current system are prepared, which include multi-label classification, and model combination between models trained on different samples of the data. Stanford also implemented two joint models that perform relation extraction and that are robust to noise in the data and the infoboxes. The models explore two different learning approaches: (1) online, without an explicit objective function, and (2) batch, using an objective function that optimizes joint prediction. Initial results are promising.

Stanford also continued improvements to their entity-linking system and participated in the TAC-KBP entity-linking task.

Stanford started development of a distantly supervised model for the reconstruction of complex event infoboxes (e.g., for terrorism events, natural disasters, etc.). While this model shares some components with Stanford’s relation-extraction system (TAC-KBP), it is fundamentally different. Unlike the slot-filling task, most of the events of interest are unnamed, which affects both the individual extractors (one can no longer extract pairs of <entity name, slot value> but rather individual slot values) and the event co-reference task, which is more complex because event fragments are discontinuous in text rather than being linked by the same co-reference chain (as in KBP). Stanford has also implemented a system that maps training infobox data to a training set of corpus exemplars. This mapping is noisy; an analysis of the extent and distribution of noise was undertaken via manual annotations.

4.5.1.3 UIUC

UIUC and CSLI collaborated in two directions: (1) developing a Textual Entailment corpus with detailed per-phenomenon annotation and (2) temporal reasoning. On (2), UIUC continued to develop capabilities in the area of temporal reasoning to augment the work on event recognition with spatiotemporal analysis. UIUC improved their baseline temporal reasoning approach that links events to temporal expressions and generates a timeline of events [132].

UIUC developed extensions to its semantic role-labeling (SRL), with the goal of extending SRL beyond verb predicates to a number of other relations. The focus is on integrating nominal relations, verb-based relations, and prepositional-based relations. UIUC investigated several global-inference approaches to support this process. One of the key difficulties in developing a SRL system that handles multiple types of relations is that no annotated source covers all relation types. UIUC developed a JI approach that can use existing structure predictors as black boxes and can enforce consistency constraints between the predictions. Without retraining the individual models, UIUC showed improvements in the performance of both tasks.

UIUC worked on improved named-entity resolution and co-reference resolution [133]. UIUC developed a new ILP formulation for co-ref, as well as several new training algorithms to explore mention-based approaches, entity-based approaches, and joint learning approaches. UIUC's co-reference system won two of the four evaluation metrics at the CoNLL-11 Shared Task Evaluation, including B3, the most commonly used evaluation metric. UIUC's average score was third (in a statistical tie with the second place team). UIUC participated in the co-reference-resolution competition for the health records domain. This system was in the top eight of the I2B2 competition and was selected to the JAMIA journal.

UIUC continued its work on Wikification: mention identification, disambiguation, and the mapping of mentions to the appropriate Wikipedia page. UIUC compared global approaches that simultaneously consider multiple mentions and take the hyperlink structure in Wikipedia into account to local methods that are more similar to traditional WSD methods. The goal is using Wikification for resolving co-reference within and across documents. UIUC is developing ways to use Wikification as a knowledge-acquisition method. UIUC augmented the Wikification capabilities with a module that assigns abstract categories to each concept of interest.

UIUC developed an improved co-reference approach that makes use of knowledge acquired by the Wikifier. UIUC explored the interplay of knowledge and structure in co-reference resolution. UIUC explored ways of using the "grounding" of mentions into Wikipedia provided by the Wikifier for boosting performance. To maximize the utility of the injected knowledge, UIUC developed a layered-learning approach that, at each layer, performs entity-level inference. This system outperforms the state-of-the-art baseline by two B3 F1 points on the non-transcript portion of the ACE 2004 dataset.

UIUC worked on a better event-identification and mention-detection approaches, to aid event extraction, co-reference, named entities, and Wikification; the emphasis was on incorporating this module in other tools in a modular way, without affecting the trained model, to support a move to a new domain.

UIUC continued work on relation recognition and on event recognition, tracking, and de-duplication. UIUC's approach emphasized an integration of the mention-detection process with the relation classification. This approach makes use of global inference over multiple components, including a Wikifier, co-reference resolution, and enforces coherency constraints among relation types.

UIUC continued work in event recognition, focusing on the identification of events and supporting events, and the identification of causality and “relatedness” relations among them. Specifically, UIUC’s approach builds on two types of information sources that contribute to identifying relations among events: First, UIUC developed a minimally supervised approach based on focused distributional similarity methods for extracting causality relations between event pairs in context. Second, based on the observation that discourse connectives and the particular discourse relations that they evoke in context provide additional information about causality between events, UIUC developed a bank of discourse-relation predictors. UIUC then showed that combining discourse-relation predictions with the distributional similarity methods in a global inference procedure provides additional improvements, culminating in the ability to recognize causality relations among identified events.

4.5.1.4 University of Massachusetts

UMass worked on jointly resolving entities of different types (people, authors, papers, and venues) with minimal supervision over Bibtex-style MongoDB records, scaling up an initial version to large quantities of Bibtex records. UMass investigated ways of incorporating human edits to this data in a probabilistic manner by treating them as evidence in a graphical model. UMass began exploring a new entity-based model that is better able to trade off the human edits with the machine predictions. UMass implemented a new, infinitely deep hierarchical model of co-reference that communicated with the MongoDB back end. Additional improvements were made to both the model and the backend, including both for speed and co-reference accuracy.

UMass worked on large-scale joint co-reference, segmentation, and alignment for semi-structured data. Their goal was to cluster records and labeled/segmented texts without any supervision. They use a conditional random field model with posterior constraints for inference/learning. Significant progress was made in terms of speed of inference. Learning was further speeded by using asynchronous online learning of CRF weights. On one dataset, a significant performance improvement was observed via joint co-reference and segmentation.

UMass implemented a unified approach to both the bootstrapping of relations and distant supervision, based on the framework of Posterior Regularization (PR). They worked on models that (1) support discovery of new relation types and entity types, and (2) applied distant-supervision and bootstrapping to the task of biomedical relation extraction. Part (a) led to the “mention completion” view stated below. In part (b), they trained a model using an existing protein-protein interaction database. Crucially, in contrast to most distantly supervised methods, no closed world assumption when learning is made.

UMass began work on a paradigm for information extraction inspired by image completion. The goal is to train models that take partial mentions of entities, or entity pairs, and complete these with information about what else one can say of the entity. Mention completion can be realized through various means, including deep networks and matrix completion. UMass learned the structure of a pairwise graphical model to complete mentions. In the simplest case, this model was required to be a tree. In this case, learning the structure amounts to finding the maximum spanning tree with respect to mutual information between features.

UMass developed improved methods for constituency parsing and its integration with named-entity recognition. By expressing a labeled bracketing model in a factor graph, state-of-the-art parsing on OntoNotes can be achieved without consulting a large grammar or, building on UMass work in 2010 on marginal relaxation, by inducing only a small number of grammar-rule constraints. Pruning techniques similar to the left-corner pruning in chart parsers were adapted for use in these factor-graph models of constituency parsing.

UMass worked on a new version of their REXA platform, a digital library and search engine for research literature. This version performs joint citation, author, and venue co-reference. Human edits are supported and integrated into the probabilistic reasoning process. Inference uses a hierarchical and entity-centric sampling approach implemented within FACTORIE. A particular focus was on developing a database schema that allows for both efficient inference and efficient navigation through the web frontend, while scaling up inference to handle all of DBPL.

UMass developed new techniques of unsupervised domain adaptation based on information-retrieval models for context aggregation and applied them to improved named-entity recognition. Using information-retrieval models addresses some difficulties in graph-based, semi-supervised learning: natural-language problems often lack well-motivated similarity functions with small numbers of parameters. This approach significantly outperformed state-of-the-art NER systems when using unlabeled data to adapt to news domains.

UMass worked in unsupervised entity-type refinement for NER using topical information. UMass assumes that documents with different topics contain different entities and, at higher degrees of granularity, different entity types. This work refined entity types via a split-merge procedure, using topic distributions to guide cluster initialization.

UMass explored large joint models of NLP with hidden syntactic variables. By constraining hidden variables to adhere to tree structure, and marginalizing out this hidden structure to optimize performance on the end task, reliance on jointly annotated data or pre-processing with trained parsers was reduced. These models may exhibit greater performance on languages whose syntactic structure is less strict.

UMass improved their biomedical event-extraction system in collaboration with Stanford [67]. They found that by intersecting the results stacking with those of taking a union of Stanford's and UMass's, further improvements could be achieved. This technique effectively removes novel events not proposed by any of the base systems.

UMass worked on unsupervised relation discovery. They investigated a two-stage approach for clustering paths and entity pairs connected by them: intra-path clustering and cross-path clustering. In the first stage, they interpreted each path by clustering its entity pairs. Each interpretation is represented by a subset of entity pairs of the path. By dividing a pattern's entity pairs into interpretations, different interpretations can have different sets of similar paths. In the second stage, they merged interpretations of different paths to get semantic relations. One path can fall into different semantic relations due to its different interpretations. They explored local features, such as the words between the two entities and global features such as the theme of the document and sentence in which the entity pair and dependency path occurs. They evaluated the clusters predicted by their approach with respect to Freebase relation clusters, and observed substantial improvements in comparison to their older work, which does not differentiate between different senses of the same path.

4.5.1.5 PARC and Stanford's CSLI

PARC and CSLI continued the implementation of its finite-state temporal annotation tool based on the interval calculations done in the PARC rule-system. The tool is meant to be run with various parsers. The aim is to annotate temporal modifiers in such way to enable temporal inferences that do not depend on calendrical anchoring (relative time anchoring). For example, from the representation of the temporal information in "Mary visited New York last summer" and "Ed has been living in New York for three years" a temporal reasoner can infer that Mary visited New York while Ed lived in New York. The appropriate representations are derived by marking the temporal reference point, a time interval, and the relation of an event to that interval.

This approach splits the annotation that is done in XFST from the normalization and normalization/semantic interpretation that is done in Python. In particular, an analysis of the contribution of various temporal prepositions was codified. Xerox's Finite State Tools (XFST) have been independently expanded to have facilities for pattern matching and context-free parsing using recursive transition networks. PARC developed XFST scripts for recognizing date patterns, and non-calendrical temporal patterns (which is an advantage over the existing GUTime system), such as temporal prepositions.

For dates, PARC used the Enron corpus to learn the contexts that distinguish dates from other non-temporal expressions (e.g., fractions, like "1/2 kilograms," from dates expressed as "1/2"—January 2). PARC evaluated their date system along the TempEval gold standard set.

PARC used the Google n-gram corpus for learning temporal prepositions patterns, and developed syntax for expressing different types of temporal prepositions along the lines of existing linguistic work. For this work, PARC started developing a gold-standard dataset, which, to their knowledge, has not been built for evaluating temporal prepositions. PARC also used the n-gram corpus to distill a list of event nouns that was distributed to the FAUST community.

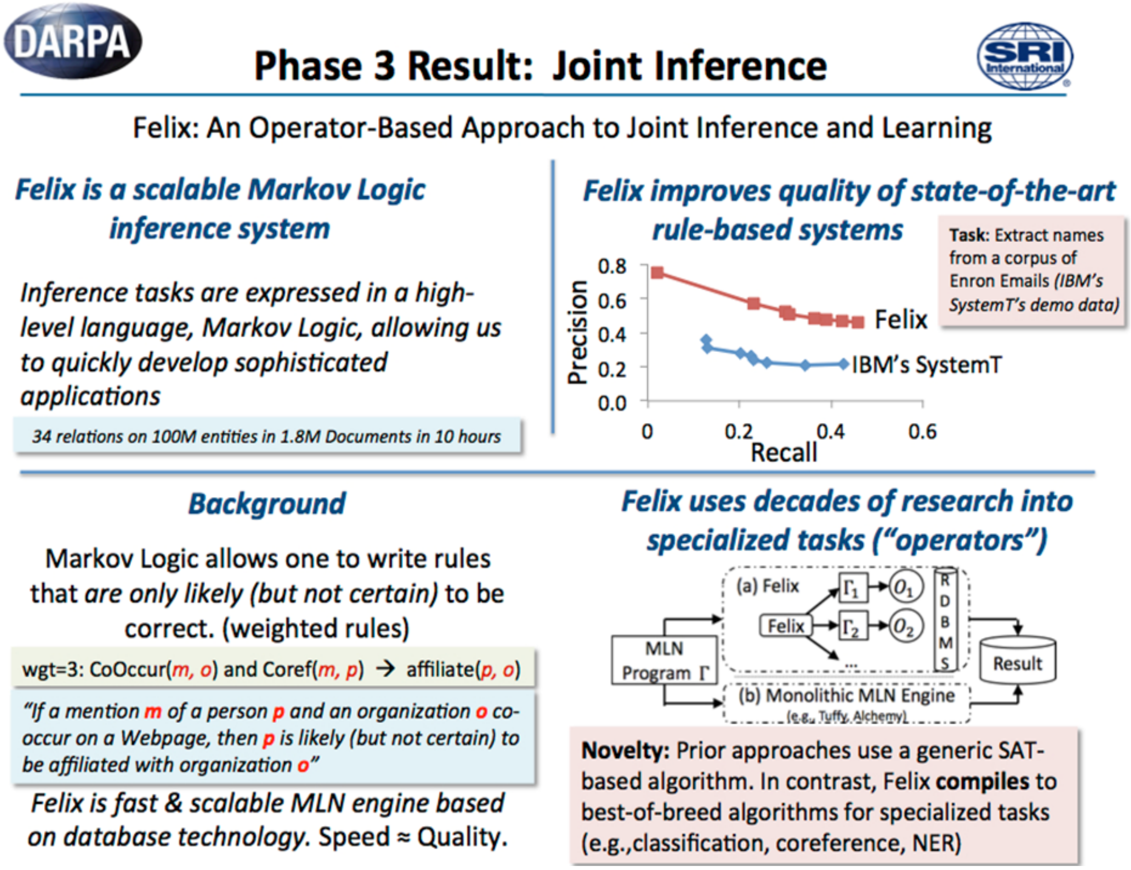
PARC worked on the NSF workshop proposal on the Semantics for Textual Inference that took place at the LSA Summer Institute at the University of Colorado in July 2011.

4.5.3 Joint Inference

4.5.3.1 SRI and Wisconsin

SRI achieved an initial implementation of Lifted Belief Propagation (LBP), including a symbolic evaluation package, an equality constraint solver and model counter, and the Lifted Belief Propagation algorithm itself. These capabilities serve as a basis for further developments including Anytime LBP and First-order Variable Elimination.

The University of Wisconsin developed the Felix system [85], which decomposes high-level statistical-inference tasks (such machine reading) into a set of low-level database tasks, as shown in Figure 8. In a performance study that compared the quality of the Felix approach to several state-of-the-art techniques including commercial offerings (e.g., System T from IBM), Wisconsin determined that Felix offers higher quality (and somewhat surprisingly) higher runtime efficiency than prior approaches for simple information-integration tasks. Felix is publicly available at http://hazy.cs.wisc.edu/hazy/felix/.



Phase 3 Result: Joint Inference

Felix: An Operator-Based Approach to Joint Inference and Learning

Felix is a scalable Markov Logic inference system

Inference tasks are expressed in a high-level language, Markov Logic, allowing us to quickly develop sophisticated applications

34 relations on 100M entities in 1.8M Documents in 10 hours

Felix improves quality of state-of-the-art rule-based systems

Task: Extract names from a corpus of Enron Emails *(IBM's SystemT's demo data)*

Background

Markov Logic allows one to write rules that *are only likely (but not certain)* to be correct. (weighted rules)

wgt=3: CoOccur(*m, o*) and Coref(*m, p*) → affiliate(*p, o*)

*"If a mention **m** of a person **p** and an organization **o** co-occur on a Webpage, then **p** is likely (but not certain) to be affiliated with organization **o**"*

Felix is fast & scalable MLN engine based on database technology. Speed ≈ Quality.

Felix uses decades of research into specialized tasks ("operators")

Novelty: Prior approaches use a generic SAT-based algorithm. In contrast, Felix **compiles** to best-of-breed algorithms for specialized tasks (e.g.,classification, coreference, NER)

Figure 8: Phase 3 results for joint inference at the University of Wisconsin

The SRI PCE team explored different ways of doing temporal reasoning and event co-reference in the IC++ and MR-KBP domains. Multiple MLNs were created incorporating knowledge from both the MR-KBP and temporal ontologies. They analyzed the Wisconsin team's output, to see if there are ways of doing collective inference, formalized the problem of learning MLN weights

from subjective probabilities, and analyzed the rules output from the UMass system for relation extraction.
Wisconsin leveraged the techniques and infrastructures that they built for MR-KBP to a demonstration system called Wisci/DeepDive. Wisci collects raw text from various sources (including a 500-million-webpage corpus called ClueWeb and hundreds of thousands of YouTube videos—both metadata and transcripts) and runs their reading system end to end. For each Wikipedia page, Wisci surfaces related mentions in Web pages and videos; Wisci also adds text and video provenance for infoboxes. This demo is running and available online at http://hazy.cs.wisc.edu/hazy/deepdive/.

Wisconsin and Wake Forest collaboratively developed MLN-BOOST, a functional gradient-boosting algorithm for MLNs. This algorithm enables rapidly and accurately learning the structure and the parameters of an MLN. Other methods for learning MLNs follow a two-step approach: first, perform a search through the space of possible clauses, and next, learn appropriate weights for these clauses. Wisconsin instead simultaneously learned both the weights and the structure of the MLN. This approach is based on functional-gradient boosting where the problem of learning MLNs is turned into a series of relational functional approximation problems. Two representations for the gradients were used: clause-based and tree-based. Their experimental evaluation on several benchmark datasets demonstrates that MLN-BOOST can learn MLNs as well or better than those found with state-of-the-art methods, but often in a fraction of the time. For further details on MLN-BOOST, see [93][8].

Wake Forest continued the implementation of Anytime Lifted Belief Propagation. Their current system scales well and can handle non-loopy graphs correctly.

Wake Forest developed an infrastructure for automatically analyzing the documents available in the TAC-KBP 2010 corpus and extracting information about the entities (primarily people and organizations) within those documents. This information includes such data as age; city of birth; subsidiaries; shareholders; website; etc. From these attributes, a set of first-order logic predicates is produced, enabling the extracted information to be easily used in other software. To this effect, they have integrated the Lucene search engine with the TAC-KBP data and have developed a UI for also allowing human annotation of some unlabeled data.

[8] http://ftp.cs.wisc.edu/machine-learning/shavlik-group/khot.icdm11.pdf

4.5.3.2 University of Washington (UW)

The University of Washington (Prof. Domingos) continued working on USPN, a new end-to-end solution to machine reading that extends USP to process text online. USPN builds on three synergistic components. The first component is sum-product networks (SPNs), which enables efficient inference by compactly representing the partition function as a deep network. The second component is probabilistic theorem proving, which provides a unified approach to joint inference by combining lifted inference and sampling. The third component is a new coarse-to-fine paradigm based on mean-field approximation.

UW started investigating a grammatical formulation of unsupervised semantic parsing and began developing a linear-time parsing algorithm for it. UW continued working on sum-product networks (SPNs) for efficient inference and presented this work in UAI and IJCAI [84].

UW (Domingos) introduced a family of deterministic, structured message-passing algorithms for efficient JI; initial experiments show that they are substantially more accurate compared to state-of-the-art when the graphical model has structural features. UW continued developing hierarchical mean field for coarse-to-fine inference and started experimenting with a prototype system. Additionally, UW started to investigate a grammatical formulation of unsupervised semantic parsing for analyzing various approximation schemes and extensions.

UW (Domingos) improved a prototype for their hierarchical mean field. This hierarchical mean field selects an approximating distribution characterized by a hierarchy of alternating mixture mean field and cluster mean field approximations. By alternating the cluster and mixture mean fields, one can approximate a complex inference problem as a hierarchy of increasingly simpler (but increasingly many) approximations.

UW (Domingos) conducted a deeper analysis of update schedules for hierarchical mean field. Some update schedules of variational parameters in a hierarchical setting can lead to unstable behavior. UW found a level-by-level approach to parameter updates can instead produce stable behavior. This tiered optimization approach, while bottom-up for now, could theoretically be adjusted to work in a top-down way, which would allow for easy integration of coarse-to-fine probabilistic inference.

UW (Domingos) developed an algorithm for coarse-to-fine variational inference based on recursively refining mean field mixture models until either an accuracy or a time bound is reached. This complements the work in the previous item, providing an inference algorithm for intractable probabilistic knowledgebases with much better approximation than was previously possible. Experiments trying to approximate distributions that are true mixture distributions show that the approach can correctly approximate these distributions and the approximation quality improves after each refining step until the true number of mixture components is reached.

UW (Domingos) continued their work on structured message-passing algorithms, strengthening their theoretical results, and performing empirical tests on large, real-world problems. Unlike existing approximate message-passing algorithms, their new algorithms exploit logical structure that is quite prevalent in real-world problems.

UW (Domingos) developed new rules for lifted importance sampling. These rules detect substantially more symmetries in the first-order representation than existing rules [42], which are designed primarily for exact inference. UW proved that these new rules are sound and reduce the variance (increase the accuracy) of lifted sampling.

UW started to develop a tractable subset of Markov Logic, in which efficient JI with specified accuracy is guaranteed. Exploiting ontological information in the form of is-a and has-a hierarchies enables this. In turn, this enables efficient JI in a substantial subset of the probabilistic knowledgebases produced by MR, as well as at successive stages of the reading process.

UW (Domingos) continued to develop an algorithm for multiple hierarchical relational clustering. This enables a coherent probabilistic knowledgebase to be induced from the raw data obtained by reading. A version of the algorithm that learns a single hierarchical clustering was implemented and evaluated.

4.5.3.3 Other Team Members

UIUC made several key steps in their JI efforts. (1) UIUC developed a general framework containing a graded spectrum of Expectation Maximization (EM) algorithms called Unified Expectation Maximization. This is a family of EM algorithms parameterized by a single parameter and that covers existing algorithms like standard EM and hard EM; constrained versions of EM, such as Constraint-Driven Learning (Chang et al., 2007); and Posterior Regularization (Ganchev et al., 2010); along with a range of new EM algorithms. (2) UIUC developed an efficient dual projected gradient-ascent algorithm that can be used for constrained inference, which generalizes several dual decomposition and Lagrange relaxation algorithms popularized recently in the NLP literature (Koo et al., 2010; Rush and Collins, 2011). (3) UIUC made significant progress developing a Lifted ILP algorithm that, in preliminary experiments, saves a significant percentage of the computation needed in constrained optimization inference.

UMass investigated using Particle Filtering as a method for inference in large-scale probabilistic databases based on factor graphs. UMass worked on their Expectation Propagation approach to JI. They combined a pairwise co-reference model with a span-based NER model and a relation model that takes into account NER information. They applied their methods to ACE 2004 data.

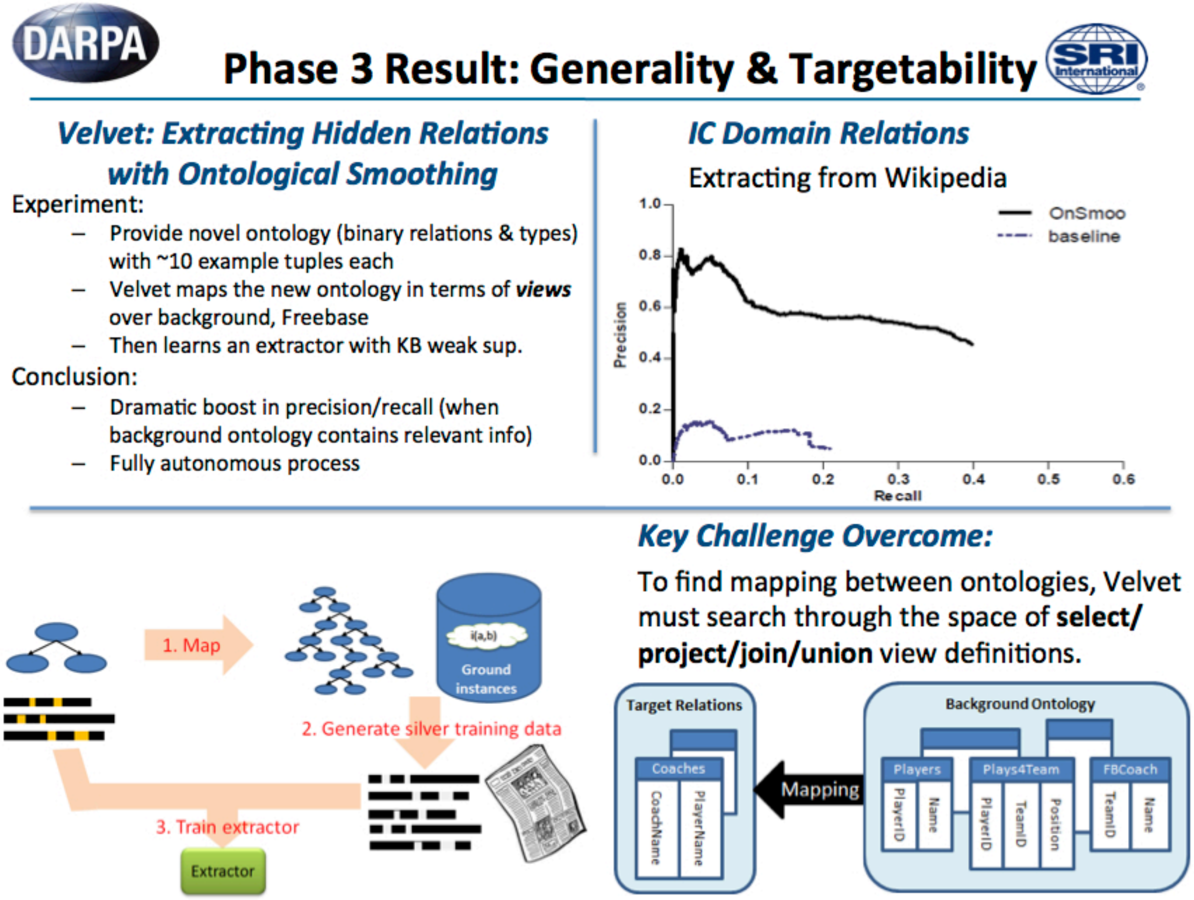
UIUC and UMass jointly developed an abstraction of a joint-inference module that will enable integrating individually learned modules and running multiple forms of joint inference for it. They improved this joint architecture, which integrates the UIUC and UMass RE modules.

Columbia continued work on algorithms for JI based on dual decomposition and Lagrangian relaxation.

4.5.4 Learning for and from Reading

The University of Washington (Prof. Weld) extended their fine-grained entity recognizer to handle a new tag-set requested by UIUC for use with the IC++ domain. They expanded their internal evaluation to handle newswire text. They experimented with domain-adaptation techniques and co-reference analysis. They started applying knowledge compilation, using an ensemble of NER systems to create silver training data. They showed that the resulting fine-grained entity recognizer created features that strongly improved the precision and recall of relational extraction. A paper on this research was published at AAAI 2012 [136].

As shown in Figure 9, UW completed their implementation of and experimentation on the Velvet ontological smoothing system, a semi-supervised technique that learns extractors for a set of minimally labeled relations. Ontological smoothing has three phases: First, it generates a mapping between the target relations and a background knowledgebase. Second, it uses distant supervision to heuristically generate new training examples for the target relations. Finally, it learns an extractor from a combination of the original and newly generated examples. Experiments on 65 relations across three target domains show that ontological smoothing can dramatically improve precision and recall, even rivaling fully supervised performance in many cases. A paper on VELVET was published at AAAI 2012 [137].



Phase 3 Result: Generality & Targetability

Velvet: Extracting Hidden Relations with Ontological Smoothing

Experiment:

- Provide novel ontology (binary relations & types) with ~10 example tuples each
- Velvet maps the new ontology in terms of ***views*** over background, Freebase
- Then learns an extractor with KB weak sup.

Conclusion:

- Dramatic boost in precision/recall (when background ontology contains relevant info)
- Fully autonomous process

IC Domain Relations

Extracting from Wikipedia

Key Challenge Overcome:

To find mapping between ontologies, Velvet must search through the space of **select/ project/join/union** view definitions.

Figure 9: Phase 3 results for information extraction at the University of Washington

UW continued development of their integrated development environment for rapid debugging of rule-based extractors. Features include: (1) the ability to interactively define new features, which are immediately incorporated into a linear classifier, (2) integrated statistics for interactively evaluating performance of the system on a development set, and (3) a logic-to-SQL compiler that enables the system to provide interactive performance even on large corpora, such as the full *New York Times* archive. The interface allows a human knowledge engineer to quickly search for keywords (e.g., "assassinated") and to see matching sentences; then to click to automatically generate a possible prolog-style rules that match that sentence (e.g., terms on the right-hand side of the rule correspond to dependency relations and lexical entries); and finally then to click to choose a candidate rule and immediately see all the extractions resulting from that rule in a large corpus (e.g., historical *New York Times*). One can then refine the rule or add additional rules and rules can chain. This supports an extremely rapid interactive development cycle for extractor development. Preliminary experiments were carried out on several relations, such as *PersonKilledByPerson* relations, and it appeared that high precision and moderate recall rule sets could be authored in approximately an hour.

Columbia University developed a new spectral learning algorithm for latent-variable PCFGs (L-PCFGs). Previous work on L-PCFGs had used the EM algorithm, which is only guaranteed to reach a local optimum of the likelihood function. The spectral algorithm is guaranteed to give consistent parameter estimates for L-PCFGs under assumptions on certain singular values that are defined in the model. The spectral algorithm is simple and efficient, relying on a singular value decomposition on the data, followed by simple matrix operations. The method involves a tensor-based view of L-PCFGs that may be useful in other contexts. Columbia's initial work in this area developed the basic algorithm, and the theory underlying it; in Phase 3R, Columbia implemented experiments with the method.

UMass worked on a distributed learning approach that allows asynchronous inference workers to send small messages consisting of local gradients that are used to update the weights at a central learner. The SampleRank update was implemented and tested on distributed training on IID instances.

UIUC developed a new, principled framework for performing efficient learning with declarative constraints over structured output spaces. This framework establishes a spectrum of learning algorithms ranging from global learning to independent models, allowing a customization of the hardness (and efficiency) of learning based on a given set of constraints. UIUC's theoretical results provide a general combinatorial characterization of an arbitrary set of constraints under which these algorithms are consistent. UIUC conducted experiments that show these algorithms are robust and, while being significantly more efficient, provide performance close to global learning.

UIUC developed a taxonomic relations classifier that makes use of stationary ontologies and Wikipedia to infer robustly is-a relations and sibling relations. The approach makes use of global inference over multiple related relations.

UIUC continued their efforts for developing better tools and a language for their Constraint Conditional Models (CCMs). The language facilitates learning-based programs with an easy way to declaratively define constraints and support global inference with them.

4.5.5 Infrastructure, Software Engineering, and Integration

Early in Phase 3, the SRI SE Team decided to restructure FAUST to remove dependencies upon any software provided by the ET, which had proved troublesome during Phase 2. We focused on the engineering problems of scaling up to 1.5 million documents and purchased the additional hardware necessary to accomplish the evaluation goals.

The SE Team invested significant effort in building up the tooling to analyze the results of subcontractor-provided modules and to ensure that the new releases did not introduce a regression as had occurred in Phase 2. We took on the task of converting from existing data formats to the newer input and output formats being developed by the ET for Phase 3.

The SE Team developed a Gazetteer module to enable the gazetteer released by the ET to be useful to other software components. This effort included augmenting the data.

The SE Team manually annotated hundreds of documents to provide test and evaluation data to facilitate testing and improvement of MR modules.

Wisconsin delivered updates and new components for the MR-KBP and TAC-KBP evaluations. These included:

- A baseline end-to-end system for MR-KBP, which takes as input the MR queries from the training data and outputs extracted assertions that can be directly evaluated. The output is a set of temporal relations between mentions. This system is built upon the non-temporal slot-filling system and employs rule-based temporal resolution that maps temporal expressions in natural language to normalized temporal intervals,
- Tools developed in collaboration with SRI to ground the bound entities and convert queries from the ET into SQL commands for interfacing with the PostgreSQL database and convert the database responses back into CAF,
- Wisconsin's annotation tool developed to annotate a set of documents to provide material to fill in gaps not covered by training examples, or where training examples were wrong with respect to the guidelines.

Wisconsin implemented Felix, a relational optimizer for statistical inference, which contains Tuffy inside. Felix implements operator-based task decomposition to identify and perform specialized subtasks of a given MLN program. Among the operators that Felix supports are classification, labeling, co-reference resolution, and MLN inference. Felix also implements *dual decomposition* for more efficient inference.

Stanford continued to update their NLP components and refactored their Stanford CoreNLP framework to include the enhancements made as part of MRP-specific work for their NFL annotator. Stanford released updated versions of the Stanford CoreNLP pipeline with:

- Improved co-reference system
- Initial version of SUTime for temporal expression identification and grounding.
- Improved tokenizer and sentence splitter to do whitespace tokenization.
- Ability to incorporate custom annotators

Stanford incorporated Semgrex, a system for matching regular expressions over dependency graphs to its Tregex package, which includes tools for matching regular expressions over parse trees and applying transforms to parse trees (Tsurgeon). This package is used extensively in co-reference and in parts of the system that use dependencies.

Stanford released a thread safe version of the Stanford parser and the Stanford Biomedical Event Parser.

UMass improved the FACTORIE platform through bug fixes and performance improvements.

UMass improved their persistence layer based on MongoDB that supports processing and storing of large-scale corpora. It uses a lightweight façade approach in which clients write typed façades that wrap around Json/Bson objects. It supports storing typed objects into databases, as well as file-based storage. This has significantly simplified working with large corpora and allows efficient distant supervision for event extraction in IC++.

UIUC deployed versions of its Curator-based NLP tools to SRI. These include the Wikifier, and seven other tools, including co-ref, and an all-Java version of their Semantic Role Labeler.

UIUC released a prototype of its web-based Event Annotation Tool (EAT), which was updated to be configurable to new ontologies, and was modified to respond to requests from SRI. SRI used it to annotate documents for IC++.

UIUC continued to improve their completely new EE (event extraction) system designed around joint-inference architecture. This modular architecture promotes flexibility by abstracting away component implementation and by providing a simple but flexible API for components performing well-defined tasks, which allows for interdependency and joint reasoning over arbitrary subsets of components.

4.5.6 Use Cases and Evaluation

The Phase 3 evaluation was redefined to include only the MR-KBP task as well as to greatly reduce the scale of the evaluation. SRI's SE team created a new system under the accelerated schedule to perform the MR-KBP evaluation.

SRI performed extensive testing and quality assurance of each release by the ET, pushed bug reports to the ET, and identified many inadequately defined evaluation parameters and formats. This effort proved invaluable, as we were able to detect inconsistencies in several of the evaluation questions. These required the ET to re-release the evaluation questions. Essentially, our SE team was performing a quality-control role for the ET.

The SE Team executed and passed the evaluation. The results of the FAUST system on the Phase 3 evaluation were quite good.

Wisconsin conducted a literature review to identify additional training data to supplement the data made available by the ET, because of the relatively small size of that corpus and inconsistencies in the training labels. They identified two relevant datasets: TimeML and TempEval-2010; the latter was chosen because of its similarity to the MR-KBP task.

- The Wisconsin team converted the TempEval dataset into a first-order dataset for use with their boosted RDN/MLN learning system. This baseline system was only 25% behind state-of-the-art approaches on various tasks over the training set even though it had never been used for any NLP task before and was not tuned for the tasks. In contrast to state-of-the-art approaches, which only learn model parameters, this system learned the model structure as well as its parameters.
- Using this baseline system, Wisconsin developed methods specifically for the temporal aspect of the slot-filling task. Progress was made on the three main aspects of the problem: identifying the verb that a time phrase refers to; identifying the relationship that the verb modifies; and finally, determining the effect (initiation, continuation or termination) of the verb on the relationship.
- Wisconsin translated patterns in the dependency graph that are indicative of the three aspects of temporal attachment into first-order predicates to be used as features in this learning system. Two approaches for temporal attachment were developed: (1) a structure-learning approach and (2) a parameter-learning approach that learns weights for handwritten MLN rules.

Wisconsin worked on the entity-linking and slot-filling tasks from TAC-KBP as a surrogate for the MR-KBP task.

- For the slot-filling task, Wisconsin investigated using the annotations provided by TAC for the 2009 and 2010 slot-filling task for training. However, this provided too few training examples for nearly all of the relations of interest, so Wisconsin used these relations as a ground-truth test bed for evaluating their algorithms. Unlike MR-KBP, only the slot fillers are marked by TAC in the document; the entities were not marked.
- Wisconsin improved their approach by systematic error analysis and implemented changes to include using a set of MLN rules to prune the training data, and generalization of patterns. They also used a combination of two features: (1) dependency path and (2) word sequences to improve the quality. Wisconsin's F-1 score on TAC-KBP 2010 slot-filling task was 34.

Wisconsin worked extensively on the MR-KBP task for Phase 3.

- Using the pilot annotations provided for the KBP task, Wisconsin converted the documents and annotations into first-order logic facts. They built a baseline system using Conditional Random Fields (CRFs) for entity detection and logistic regression for relation extraction.
- To reduce the engineering effort on feature engineering and cross validation tasks, the Wisconsin team constructed an additional component of Felix (their MLN system implemented using an RDBMS) to support high-throughput feature engineering, learning of weights, and cross validation.
- To handle cases where examples are not sufficiently numerous, Wisconsin used Freebase to obtain distantly supervised examples. For all relevant entity pairs in the Freebase relations, such as *spouseOf*, the sentences containing both the entities are considered as positive training examples. This provides a large set of training examples although they are rather noisy. In addition, Wisconsin collaborated with Prof. Weld’s group at UW, who provided additional distantly labeled examples from Wikipedia and TAC-KBP.
- Wisconsin added features into their temporal slot-filling system. In particular, to leverage redundancies in text, they expanded the document set with results from Google queries and ran their slot-filling system on the expanded document set. Wisconsin developed a more sophisticated temporal-axiom system and a temporal-attachment system to improve quality on temporal slot filling. With these additional features, Wisconsin was able to boost their F1 score on the 75-document training corpus of MR-KBP to 40.
- Wisconsin continued their work on using domain knowledge to extract NFL game descriptions.
- They added high-precision, handwritten rules to extract NFL relations from sentences such as “Packers 31 Bears 27.” This improved the precision on their system evaluated over a database of NFL games that contains the date, winner, loser, home and away teams, and final score for each game.
- Wisconsin worked on using Stanford's temporal extractions to generate all time mentions. They converted the time expressions generated by Stanford to a range of dates. This enabled resolving the conflicts between various temporal extractions. Wisconsin worked on writing inference rules for resolving such conflicts.

Wisconsin developed visualization tools to help with the annotation and evaluation aspects of the various datasets.

- Wisconsin developed a web-based GUI for collecting additional annotations to augment those provided by the ET. They tested it and made it available to other SRI team members.
- Wisconsin developed a visualization tool to evaluate the results of their end-to-end system. This tool allows team members to record notes on wrong answers provided by the system, as well as to convert these answers into additional training examples for future runs.

Stanford's deterministic sieve co-reference system was the top-ranked system at the CONLL-2011 shared task on co-reference resolution over OntoNotes English data, beating out all machine-learning-based systems. Stanford released this updated version of the co-reference system as part of the Stanford CoreNLP distribution.

Stanford again participated in the 2011 NIST TAC KBP Entity Linking and Slot Filling tasks.

UIUC developed a complete Event Extraction for IC++ based on their EMNLP work, enhancing the event identification, expanding the coverage of event types and argument types. In particular, the system also detects relations (causality) between relations, taking into account both discourse connectives and distributional similarity. Significant progress was made on:

- Identifying and parsing events: improving the unsupervised method for identifying events and parsing them to identify arguments.
- Temporal Reasoning: extracting temporal phrases, mapping temporal phrases to recognized events, and generating a time line of events.
- EAT was improved based on feedback from SRI and renamed as EAT+.

UIUC improved its temporal-reasoning component. UIUC currently only deals with temporal reasoning with respect to extracted temporal phrases for six types of relations (before, after, overlaps, etc.). The approach also deals with time-lining events.

UMass used distant supervision for IC++ events. They use temporal knowledge about existing events (such as the 2005 New York City mayoral elections) to select newswire articles. These articles are then scanned for event arguments and candidate trigger words, and the assumption is made that mentions of the arguments tend to correspond to mentions of the events. Currently, trigger words are hand-specified, but UMass is working on automatically generating these triggers from data. A baseline for learning trigger words from distantly matched entities was implemented. The baseline predicts trigger words according to a multi-variate Bernoulli model, incorporating NER, POS, and parse tree information.

4.6 PHASE 3-RESEARCH RESULTS

In Phase 3R, we report our results by institution, as each institution was conducting its own DARPA-approved research program. DARPA gave explicit guidance to SRI to not use resources to produce an end-to-end system in this phase and to instead concentrate on the research goals.

4.6.1 Stanford University (Prof. Manning)

Stanford released several new versions of its software distributions and prepared a streamlined version, which is usable without the need of any external libraries (which have their own licenses). Stanford addressed bug reports for the new thread-safe parser. Stanford created new case-less versions of their parser and part-of-speech taggers. Stanford released a new version of SUTime that is configurable using modifiable rules for how to map text to temporal expressions. Stanford released TokensRegex for matching regular expressions over tokens. Stanford released code for its probabilistic time parsing system. Stanford improved the performance of the Stanford POS tagger and the transformation from constituency trees to dependency trees.

Stanford completed an initial implementation of a co-reference system that jointly models entities and events. Most co-reference resolution systems focus on entities and tacitly assume a correspondence between entities and noun phrases (NPs). Focusing on NPs is a way to restrict the challenging problem of co-reference resolution, but misses co-reference relations like the one between *hanged* and his *suicide* in (1), and between *placed* and *put* in (2).

1. (a) One of the key suspected Mafia bosses arrested yesterday has *hanged* himself.

   (b) Police said Lo Presti had committed *suicide*.
2. (a) The New Orleans Saints *placed* Reggie Bush on the injured list on Wednesday.

   (b) Saints *put* Bush on I.R.

Stanford introduced a novel co-reference resolution system that jointly models entities and events. At the core is an iterative algorithm that cautiously constructs clusters of entity and event mentions using linear regression to model cluster-merge operations. Importantly, the model allows information to flow between clusters of both types through features that model context using semantic role dependencies. As clusters are built, information flows between entity and event clusters through features that model semantic role dependencies. The system handles nominal and verbal events as well as entities, and the joint formulation allows information from event co-reference to help entity co-reference and vice versa. In a cross-document domain with comparable documents, joint co-reference resolution performs significantly better (more than three CoNLL F1 points) than two strong baselines that resolve entities and events separately.

This work demonstrates that an approach that jointly models entities and events is better for cross-document co-reference resolution. However, the model can be improved. For example, document clustering and co-reference resolution can be solved jointly, which Stanford expects would improve both tasks. Further, the iterative co-reference resolution procedure could be modified to account for mention ordering and distance, which would enable including pronominal resolution in the joint model, rather than addressing it with a separate deterministic sieve.

The system is tailored for cross-document co-reference resolution on a corpus that contains news articles that repeatedly report on a smaller number of topics. This makes it particularly suitable for real-world applications such as multi-document summarization and cross-document information extraction.

Stanford annotated a corpus with co-reference relations for both entities and events. Stanford is publicly releasing this corpus in the hope that it will foster new research in this novel direction. This work was presented at EMNLP-CONLL 2012 [126].

Stanford continued to explore how to improve the performance of entity co-reference. One approach focused on improving the recall of co-reference resolution systems by exploring how to detect co-reference relations between mentions with very different words (e.g., *Google* and *the search giant*). Stanford extracted a dictionary of lexically different co-referent mentions from an English comparable corpus of tech news. The key intuition was to restrict distributional semantics to documents about the same event, thus promoting referential match. The resulting dictionary contains around 200,000 co-referent pairs and improves the performance of the Stanford co-reference resolution system by 1% on the F1 score across all the co-reference evaluation measures.

A second line of research for improving the performance of co-reference has been on identifying singleton mentions (i.e., entities that are mentioned only once and thus are never co-referred). Because few discourse entities are mentioned more than once, filtering out singletons can be helpful not only for co-reference resolution but also for other NLP applications such as protagonist identification. Stanford built a logistic regression model to identify singletons, drawing on linguistic insights about how discourse entity lifespans are affected by syntactic and semantic features. The model identifies singletons with 78% accuracy and incorporating it into the Stanford co-reference resolution system yielded a significant improvement.

Stanford continued exploring the use of deep learning (neural network) methods for NLP. Stanford developed and improved techniques for modeling compositionality in semantic vector spaces. Stanford had developed a new deep architecture that is much more powerful and strictly more general that all previous models. The principal idea is that words and phrases are represented by both a vector and a matrix that models how a word can actively alter the meaning of neighboring words.

Stanford had introduced a recursive neural network (RNN) model that learns compositional vector representations for phrases and sentences of arbitrary syntactic type and length. Stanford has extended the model to assign a vector and a matrix to every node in a parse tree: the vector captures the inherent meaning of the constituent, while the matrix captures how it changes the meaning of neighboring words or phrases. For example, the matrix of "*unbelievably*" modifies the vector of a neighbor word "*awesome*." This novel matrix-vector RNN (MV-RNN) greatly enhances hierarchical deep-learning models in terms of compositional expressiveness and is, in fact, strictly more general than all previous models. This work was presented at EMNLP-CONLL 2012 [124].

- Stanford demonstrated that the MV-RNN model could learn the meaning of operators in propositional logic and natural language. The model obtains state-of-the-art performances on three different experiments: (1) predicting fine-grained sentiment distributions of adverb-adjective pairs (e.g., *fairly/not/unbelievably awesome)*; (2) classifying sentiment labels of movie reviews (e.g., "*The film is bright and flashy in all the right ways*" as a positive review); and (3) classifying semantic relationships such as cause-effect or entity-destination between nouns using the syntactic path between them (e.g., "*The accident has spread [oil]entity into the [ocean]destination.*").
- Stanford combined this MV-RNN with traditional NLP features and obtained state-of-the-art performance on the new challenging task of classifying semantic relationships between nouns, such as part-whole or message-topic (see SemEval 2010, task 8).

Stanford improved the training procedure for the above mentioned Matrix-Vector Recursive Neural Net (MV-RNN). Due to the model's complexity, it is harder to train, so Stanford explored curriculum-learning strategies. Stanford also studied the performance of different RNN-based architectures to understand the compositionality of sentiment. In particular, Stanford implemented a tensor-based recursive neural network (T-RNN) that reaches almost the performance of the Matrix-Vector RNN with many fewer parameters. Both methods outperform all prior methods on predicting graded sentiment for all nodes of a parse tree.

Stanford considered imbuing the deep-learning models with a notion of word relatedness by means of unsupervised morphological analysis. A pre-trained word embedding is accepted as input and is improved upon by first morphologically segmenting words (e.g., *unPRE careSTM fulSUF lySUF).* A natural special case of the MV-RNN model, where stems are vectors and affixes are matrices, is then used to relate words with common stems. Preliminary results show better correlation with human judgment in measuring word similarities.

Stanford started building multi-modal models for relating images and sentences using generative recursive models. Stanford used tools developed for finding structure in language to find structure in 3D objects to get state-of-the-art performance on the task of classifying household objects. This work was presented in a paper at NIPS 2012 about Convolutional-Recursive Deep Learning for 3D Object Classification [149].

Stanford started work on a reimplementation of RNNs that can be integrated into the Stanford parser. Stanford hopes to improve the performance of the parser by integrating the semantic vectors learned by RNNs into a reranker for the parser.

To add temporal information to the Stanford CoreNLP pipeline, Stanford created SUTime, a Java library that recognizes and normalizes temporal expressions using deterministic patterns. Stanford extended this package by providing a general language for specifying rules. New rules for recognizing temporal expressions can be added using regular expressions over tokens (a separate component TokensRegex was developed for easily matching regular expressions over tokens). SUTime was also extended to support text expressions involving holidays. For instance, SUTime can recognize that "Thanksgiving, 2012" refers to 2012-11-22. SUTime is similar in functionality to the Perl GUTime library. Stanford provided access to a servlet that shows the results of the SUTime and GUTime. Evaluation on TempEval2 shows that SUTime achieves state-of-the-art performance, outperforming GUTime. Information about SUTime (including an online demo) is available at http://nlp.stanford.edu/software/sutime.shtml. SUTime is distributed as part of the Stanford CoreNLP pipeline. SUTime was presented at LREC 2012 [101].

Stanford created a probabilistic parsing system for identifying and grounding time expressions into a representation compatible with the temporal slot-filling task. They aimed to expand on the range of expressions that are handled by regular-expression-based matching methods, to more elegantly handle ambiguity in the lexicon (e.g., “last week” vs. “last week of May”), and to allow for training from an arbitrary time-expression tagged corpus. This grounding is learned in a loosely supervised setting, trained on <phrase, grounded time> pairs while inferring the latent compositional structure. Unlike GUTime and SUTime, this system provides distributions over possible groundings. The system for interpreting temporal expressions was successfully combined with a CRF for detecting temporal expression. This work was presented at EMNLP 2012 [125].

Stanford completed a project on constructing timelines of significant events for interrelated entities using recent techniques from JI. The approach is to encourage agreement between two distantly supervised models. The first model performs temporal extraction to map events to their time spans using textual evidence. The second model measures consistency between events and allows learning of constraints and tendencies. Examples of these include “people typically go to school when they're 6–21 years old,” “people typically get married after attending at least one school,” “children are born when their parents are 20–40 years old,” and “people can't work at an organization until it has been founded.” By learning these patterns, the resulting timelines should be more consistent. Joint inference combining these two models can be performed with Gibbs sampling. Ultimately, these models can be evaluated in the Temporal KBP or MR-KBP-style tasks. This work was presented at EMNLP-CONLL 2012.

Stanford worked on a distantly supervised model for the reconstruction of complex event infoboxes (e.g., for terrorism events, natural disasters, etc.). While this shares some components with Stanford’s relation-extraction system (TAC-KBP), it is fundamentally different. Unlike the slot-filling task, most of the events of interest are unnamed, which affects the individual extractors (one can no longer extract pairs of <entity name, slot value> but rather individual slot values) and the event co-reference task, which is more complex because event fragments are discontinuous in text rather than being linked by the same co-reference chain (as in KBP). Stanford continued to implement and test event-detection systems within the infobox-filling task, exploring methods for learning event infobox completion from noisy, auto-generated training data. In particular, Stanford has been exploring joint methods (sentence and entity classifiers are modeled jointly) in the SEARN framework as well as probabilistic graphical modeling.

Stanford completed a novel model for relation extraction (RE) based on distant supervision (gathering training data by aligning a database of facts with text), which is an efficient approach to scale RE to thousands of different relations but introduces a challenging learning scenario. Each pair of entities from the database typically has multiple instances in text and may have multiple labels by participating in different relations. Due to this, the assignment of labels to individual instances is unknown. Further, because the alignment heuristics are not perfect, many instances do not actually express the corresponding relations. For example, a sentence containing the entities *Balzac* and *France* may express *BornIn* or *Died*, an unknown relation, or no relation at all. Because of this, traditional supervised learning, which assumes that each example is explicitly mapped to a label, is not appropriate. Stanford implemented a novel approach to multi-instance, multi-label learning for RE, which jointly models all the instances of a pair of entities in text and all their labels using a graphical model with latent variables. The new Stanford model performs competitively on two difficult domains, outperforming three models that were the previous state of the art. This work was presented at EMNLP-CONLL 2012 [127].

Stanford is working on learning truth values for arbitrary predicates for a probabilistic database. In particular, for any predicate consisting of a relation and two arguments, the system will output (1) a truth value of either "True" or "False" and (2) a confidence in the validity of the truth-value output. The approach is to compare a candidate predicate with predicates known to be true, collected from ReVerb and Ollie outputs over ClueWeb and Wikipedia, respectively. In particular, predicates are searched for which share the relation and one of the arguments, and a classifier is trained on similarity measures between the known argument and the candidate argument. Positive examples are provided by taking known true predicates, holding them out from the knowledgebase, and predicting their truth. Negative examples are provided by taking a known true predicate, changing either an argument or the relation, and assuming that it is false. Evaluation is done on new unseen tuples; eventually, integration with a relevant task (such as co-reference) will be implemented to show the approach's use in practical scenarios. A preliminary baseline system was implemented, achieving 70% accuracy (chance is 50%) on classifying predicates as true or false.

4.6.2 Stanford University, CSLI

Prof. Peters. To what does the choice of words and constructions of a text commit its author, explicitly or implicitly? CSLI's work on veridicality inferences is focused on determining how the information encoded in lexical veridicality signatures propagates to larger structures, yielding implications of whole sentences about their author's and other mentioned agents' commitments regarding the occurrence/nonoccurrence of events mentioned in a text. CSLI's investigation has revealed that these implications can be systematically decomposed in a way that allows them to be reliably computed automatically.

CSLI factored the calculation of veridicality inferences into three successive stages, each computing one of three interacting components: polarity, certainty modulation, and source commitment. Polarity calculation computes polarity domains, and explicitly marks whether the text within a polarity domain presents an event as having occurred, or as not having occurred. Certainty modulation calculates within each polarity domain the degree of certainty expressed by the text concerning the (non-)occurrence of events mentioned in that domain. Source commitment completes the calculation of events' veridicality status by indicating explicitly which agent is committed to the (non-)occurrence of which mentioned event and with what degree of certainty. Viewed from this broad perspective, veridicality inferences are the result of meaning computation, sometimes with factors that are pragmatically determined. The meaning computation propagates lexically providing veridicality information through grammatically determined domains, systematically transforming it along the way. Many veridicality inferences are fully determined by this computation. The disparate pragmatic influences that complete the underdetermined inferences depend on the surrounding text in ways that are illustrated below.

The starting point of veridicality computation is the veridicality signatures of lexical items or phrasal constructions. In this view, lexical signatures are not inert taxonomic markers but rather instructions about how a lexical item interacts with its environment—in the case of interest here, how it determines the veridicality of events mentioned in its scope and, ultimately, as signals/instructions to the reader on what inferences to draw regarding what (might have) happened or not. For example, both "manage" and "fail" embed as their grammatical complement an infinitival clause, describing an event. They both are "implicative" verbs, but under positive polarity "fail" instructs the reader to interpret this event as not really having taken place, whereas "manage" gives the opposite instruction.

Seeing annotations thus as instructions, one must construct the right ones. CSLI's approach is to enquire how the relevant environments are computed. For the example of "manage" and "fail," there are actually two relevant environments: one determining whether the verbs themselves are under positive or negative polarity (determined from structure above the verb), and the other whose polarity is reversed or preserved (determined from structure under the verb). These environments are traditionally characterized in terms of semantic representations, logical forms of some sort; but CSLI has come to see that the leaner information contained in a dependency grammar or a PCFG is sufficient. This is enough because, unlike the situations with co-reference or source reliability, lexical veridicality instructions must resort only to information that is available locally in structures provided by run-of-the-mill dependency or PCFG parsers. Specifically, lexical veridicality signatures are sensitive to, or act upon, information that is at the same clause level or one clause below, or else can be calculated via a chain of connected clauses. Moreover, the relevant information/items are in structurally predictable positions.

The three factors in calculating veridicality do not work exactly alike. Polarity calculations are sensitive to clausal (and NP-level) embedding, so a verb like "fail" presents the event in the immediately embedded clause as not having happened. On the other hand, the relation between a source and an event of reporting is generally expressed through a subcategorized participant of the reporting predicate and an embedded clause. For example, in "The spokesman said that the company had broken the rules," "the spokesman" is the subject of the reporting verb and the "that" clause is the object. The two must be treated separately so that in sentences like "The spokesman denied that the company had broken the rules" the polarity calculation follows the same pattern as with "fail," marking the rule-breaking event as non-occurring, and the relation between source and this non-occurring event is calculated separately. In "The spokesman said that the company possibly had broken the rules," the third factor comes into play: certainty modulation. This is sensitive to adjunct or other modification and in fact is cashed out at a higher level: the source is committed to the judgment that rule breaking is a possibility. On a conceptual level, keeping these aspects apart for clarity and correctness of analysis is useful. It is advisable to keep these aspects apart during implementation as well.

Propagation of these factors in computing veridicality inferences generally follows syntactic dependency structure straightforwardly, but limited complications arise from the interference of certain uses of special classes of lexical items with syntactic dependency. Neg-raising, for example, disturbs the embedding pattern of polarity calculation. Predicative adjective constructions and modal verbs disturb the pattern of certainty modulations. The parenthetical use of verbs of saying disturbs the source-event relation calculation. These, however, can be dealt with through tightly circumscribed adjustments, and do not force us to go beyond the local domains of dependency or PCFG parsing. Consequently, calculating veridicality inferences need not require recourse to semantic representations.

Discussed next is the matter of empirical validation of claims, and predictions, about particular veridicality inferences. The need for such inferences to be judged by multiple speakers of a language is widely acknowledged, because these judgments are sensitive to multiple lexical, structural, and pragmatic factors, and may be easily biased. CSLI found it useful to have examples judged not just by one small team of language experts but rather by larger numbers of competent speakers of a language. Accordingly, CSLI designed tasks in which ordinary people can participate via the Amazon Mechanical Turk to read a short passage and then render their judgment about whether an event mentioned in the passage certainly occurred, probably occurred, probably did not occur, definitely did not occur, or the passage does not warrant any of these inferences. This paradigm can support conclusions such as that one text justifies a veridicality inference with high statistical significance, while a variant of that text justifies a different inference, or ambiguously justifies the one or the other inference—but not either one deterministically. CSLI used this paradigm both to test the reliability of their own intuitions about data that was mined from open sources, such as the web, and to assess predictions that follow from their hypotheses about lexical signatures and how these signatures combine to generate veridicality inferences.

CSLI conducted experiments through the Amazon Mechanical Turk to investigate the structural and contextual features that affect the inferences that readers draw about the status of the events mentioned in a text. Items that are particularly important in this respect are those that can yield contradictory implications, which get resolved in context.

One study tested CSLI's analysis of inferences drawn from the adjective "lucky" in the syntactic frame "be lucky to X." One interpretation of "lucky" in sentences such as "She was lucky to be the third employee of Facebook" implies that being an early Facebook employee is beneficial, that she was one, and that this resulted partly from chance—it was not a sure thing. A different interpretation can be seen in "He will be lucky to keep his job after that blunder", which on one reading warrants the inference that he probably will not keep his job. This interpretation differs from the first in that respect, though they share the implication that keeping one's job is beneficial and that doing so is not guaranteed. In general then, depending on structural and contextual features, "A be lucky to X" is understood as implying `A X' or `It is unlikely that A X'. The implication of unlikelihood is associated only with the future tense in a clause without negation. Beyond those two, the structural feature that always disambiguates toward the unlikelihood implication is the presence within X of a negative polarity item (words like "any" or "ever," which require negation or more generally a downward monotone environment). Context favors one interpretation over the other by providing clues as to the utility of the outcome described by X for A and the probability of the outcome described by X. For example, contextual features that signal a high utility of the outcome described by X for A, or a high probability for X disfavors the unlikelihood implication and favors the standard implication.

CSLI conducted Mechanical Turk experiments to tease out the veridicality status of the clausal complements of adjectives. In one experiment, CSLI tested the implicative versus factive status of ambiguous adjectives, such as "smart," "brave," and "stupid," and found that for certain speakers they can indeed be implicative. In a second experiment, CSLI looked at the factors that influence the factive interpretation of constructions based on the schema 'It BE ADJ (for/of NP) to VP'. CSLI found that in the present tense, they tend to be interpreted as generics, whereas in the past, they get more easily a factive interpretation. The results need to be analyzed further to evaluate the importance of the presence of the prepositional phrase "for/of NP." CSLI's hypothesis is that the presence of the prepositional phrase and the definiteness/specificity of the NP make the construction more implicative than it is with a bare adjective.

CSLI conducted Mechanical Turk experiments to determine the veridicality status of the complement of causal predicates like "enable" and "allow." These predicates are implicative in the past tense but not necessarily in the future tense. For example, "The technology enabled the analysts to run forecasting tests" implies that the analysts ran forecasting tests, whereas "The technology will enable the analysts to run forecasting tests" implies only that the analysts may run forecasting tests and that doing so is dependent upon further conditions being satisfied. CSLI investigated the role contextual features like “awareness of choice,” “decision bias,” or “inevitability” play in favoring or disfavoring the implicative behavior of the predicate.

CSLI believes this paradigm provides a highly effective, low-cost empirical test of hypotheses and predictions about veridicality inferences, far more reliably than having them annotated by a limited set of expert judges. It also plays a valuable role in investigating the range of pragmatic influences that interact with rule-governed projection of veridicality inferences when lexical items open the door to pragmatic influence. Being able to confirm which inference is pragmatically favored in a large class of textual contexts provides a foundation for machine learning of latent or overt features of these contexts that favor one inference vs. others. This removes a major obstacle to completing the rule-determined computation of veridicality inferences from lexical signatures by also computing the pragmatic influence that context exerts.

CSLI’s work on understanding how projected lexical signatures are shaped by structural and pragmatic influences to become inferences about the status of the events mentioned in a text led us to the conclusion that the simple categorization model of lexical semantic signatures, which sees them as analogous to part-of-speech tagging or syntactic subcategorization frame assignment, is inadequate and that much more fine-grained information is needed than is traditionally recognized. This complicates the status of semantic signatures: the lexical categorization of a predicate as, say, “factive” or “implicative” can only be conditional, and any such classification needs to be done not only with specifications about the constructions the predicate occurs in but with semantic and pragmatic conditions. The consequence of that for learning lexical semantic signatures from semantically annotated corpora, such as Factbank, is that such corpora need to be orders of magnitude bigger to ensure the discovery of the factors involved in semantic signatures. While it is certainly possible to make shortcuts and use impoverished semantic signatures in certain applications, an adequate approach to lexical inference requires more targeted corpus annotation and both detailed distributional and theoretical studies of the behavior of the lexical items that warrant inferences.

Lexical Resources: Enabling automated NL understanding requires resources to assess whether events mentioned in a text are actual or not, according to the author of the text or another agent, as well as the degree of certainty of such inferences. Toward this end, CSLI compiled richer and more extensive lexicons than are generally available.

Specifically, CSLI compiled a lexicon of verbs taking infinitival complements and annotated those marked as implicative or factive for the relative temporal reference of the infinitival complement. Depending on the embedding predicate, the temporal reference of the infinitival complement may be the same as that of the embedding predicate or it may be forward shifted. When forward shifted, the temporal reference of the infinitival complement may still be more restricted by default under certain predicates but not others.

This is of importance to veridicality, as the tense of the embedding predicate and the relative temporal reference of the complement determine the factuality of the event described by the complement. For example, the predicates “manage,” “make,” and “prevail on” all imply the truth of their complement in a positive environment but differ in the constraints that they impose on the temporal reference of their infinitival complement. “Mary managed to preside over the meeting” and “Ed made Mary preside over the meeting” imply that the meeting has taken place and that Mary presided over it. On the other hand, “Ed prevailed on Mary to preside over the meeting” is consistent with the meeting occurring in the future, in which case Mary's presiding over it is expected to happen. Although the temporal reference of "to preside over the meeting" is invariably implied to be in the past when it is a complement of "managed," this restriction is lifted under certain conditions when it is a complement of "made." For example, if the means by which the causal effect is achieved is specified, the temporal reference of the infinitival complement of predicates like "make" can be in the future, as seen with "Ed made Mary preside over the meeting tomorrow by promising her a big raise."

CSLI compiled a list of phrasal implicatives, providing templates of verb-noun collocations with an implicative signature. The templates expand to more than 1000 implicative verb-noun collocations. The signature of a phrasal implicative depends on both the semantic type of the verb and the semantic type of the noun in a systematic way. For example, the implication signature for the collocations generated from the template "WASTE ASSET" is pp|nn, whereas the implication signature for the collocations generated from the template "WASTE OPPORTUNITY" (same type of verb but different type of noun) is pn|np and that for the collocations generated from the template "USE ASSET" (same type of noun, different type of verb) is pn. The equivalence classes of verbs and nouns that determine those semantic types contain items that are not together in any WordNet Synset class. For example, "acquit" and "meet," though totally unrelated in WordNet, are interchangeable in sentences such as "He conscientiously acquitted his duty to inform and educate the Court" and "The officer met his duty to investigate and had probable cause to arrest Kim." They are part of the same equivalence class of verbs instantiating the implicative template "MEET OBLIGATION pp|nn."

CSLI compiled a lexicon of adjectives taking sentential complements and annotated them for their veridicality properties following the same scheme that they developed for verbs. CSLI delivered a lexicon for two syntactic classes taking "that" clauses of such adjectives. The classes are illustrated by "John is happy that he got selected" (close to 300 adjectives) and by "It is remarkable that John got selected" (about 700 adjectives). For both classes, the semantic annotation in the majority of the cases in "factive" (i.e., the "that" clause is presupposed, hence inferred to be factual when the adjectival clause is negative or interrogative, as well as when it is positive). However, for both types there are also subsets of adjectives that do not follow this pattern and for which the annotation classes need to be extended. One well-known class is those expressing degrees of (un)certainty about the event described in the embedded clause as in "It is certain/possible/probable/uncertain/ that John was selected."

CSLI compiled a veridicality lexicon of adjectives with infinitival complements. CSLI noted that a class of such adjectives that are traditionally classified as factive are ambiguously construed as factives or implicatives. The difference is seen with negation: depending on context, "A not be stupid/brave/smart to X" can imply that `A X' or that `A not X'. For instance, "stupid" is construed factively in "He is an engineer now, he was not stupid to have studied five years to become one", which implies that he studied five years to become an engineer. By contrast, "stupid" is construed implicatively in "Itachi, being a genius and having immense intelligence, was not stupid to choose his own clan over the village," which implies that Itachi did not choose his own clan over the village. Again, context favors one interpretation over another by providing clues about the probability that X is true, or as to whether, by choosing to X or not, agent A made a good or bad decision.

4.6.4 University of Massachusetts at Amherst

Profs. McCallum and Smith. UMass reimplemented the FACTORIE platform infrastructure to allow for transparently parallel and distributed algorithms. The approach decomposes the process of parameter estimation into three parts: (1) a set of training examples, which can compute values and gradients; (2) a training strategy, which knows which examples to evaluate (and in which threads / machines); and (3) an optimizer, which knows how to take gradients and update the model's parameters. Extending this interface with new examples, trainers, and optimizers is easy, and a side-effect is that all kinds of learning in FACTORIE can be transparently parallelized in both online and batch settings, with different synchronization strategies, each being more appropriate in different scenarios.

UMass carefully optimized FACTORIE's linear algebra package, allowing for faster sparse tensors, and the gradient optimizers in FACTORIE now work with specialized linear algebra code whenever possible. UMass implemented a new generic model for linear-chain conditional random fields, with its own efficient inference algorithms, to replace the specialized implementations for part-of-speech tagging and named-entity recognition. UMass also implemented a state-of-the-art, transition-based dependency parser, which allows for fast linear-time syntactic analysis of text data. Finally, UMass improved the documentation with new tutorials and examples showcasing what FACTORIE can do and how it works.

UMass worked on a paradigm for data integration, information extraction, and alignment between structured and unstructured data sources. In data integration, UMass transformed information from a source into a target schema. A general problem is a loss of fidelity and coverage: the source expresses more knowledge than can fit into the target schema, or knowledge that is hard to fit into any schema at all. This problem is taken to the extreme in IE, where the source is natural language—the most expressive form of knowledge representation. UMass investigated probabilistic databases of universal schema, which is simply the union of all source schema. The probabilistic database learns how to predict the cells of each source relation in this union. For example, the database could store Freebase relations and relations that correspond to NL surface patterns. The database would also learn to predict what Freebase relations hold true based on what surface patterns appear, and vice versa. UMass investigated an analogy between such databases and collaborative filtering models, and used it to implement the proposed paradigm with probabilistic PCA—a scalable and effective collaborative filtering method.

UMass observed that one of computational bottlenecks of their model stems from a large number of randomly sampled negative facts. Moreover, they found that often these randomly sampled "negative" entries were in fact positive, and induced a bias that lead to prediction errors. To overcome this problem, UMass leveraged insights from collaborative filtering with implicit feedback (and without explicit negative ratings of items). One such insight is that often the only requirement is a ranking of items, rather than accurately predicting a preference score. UMass adapted this view to relation extraction: often one is primarily interested in a ranking list of extracted facts. In fact, this is how much recent work in (weakly supervised) relation extraction has been evaluated. The ranking principle can be phrased as a pairwise objective: find parameters such that all observed facts have higher score than non-observed facts. This roughly amounts to what is referred to as Bayesian Personalized Ranking in the Collaborative Filtering literature. UMass trained several relation extraction models. The first amounts to an array of classifiers, one for each possible relation. The second uses latent low-dimensional representations of both relations and tuples to rank facts. The third combines per-entity representations with per-argument-slot representations. In addition, they investigated linear combinations of these models.

UMass evaluated their approach on a large set of Freebase relations, and a set of "pattern" relations such as "X-is-the-head-of-Y." They compared their models against state-of-the-art distant supervision methods, as well as models that incorporate semantic clusters as features. To measure the quality of a ranking, UMass used well-known methods from the information retrieval literature. In particular, they calculated, for each relation, a precision curve based on manually annotated pooled results from all systems. This gave rise to an average precision per relation, and a mean average precision across relations. UMass showed that while the previous state-of-the-art method outperforms their basic classifier model, all models that induce and leverage latent representations substantially outperform state-of-the-art. In particular, they observed more than 10% point improvements in mean average precision over Surdeanu's 2012 work. They showed that combining per-tuple and per-entity models improves accuracy further.

UMass developed improved methods for constituency parsing and its integration with NER. By expressing a labeled bracketing model in a factor graph, state-of-the-art parsing on OntoNotes can be achieved without consulting a large grammar or, building on UMass work in 2010 on marginal relaxation, by inducing only a small number of grammar-rule constraints. Pruning techniques similar to the left-corner pruning in chart parsers are being adapted for use in these factor-graph models of constituency parsing. This model outperforms the Stanford and Berkeley parsers on the task of NP-span identification while being asymptotically faster. To further improve performance, a small number of rules can be learned using perceptron weight updates. Results were presented at the COLING conference in December 2012 [152].

UMass continued to develop new techniques of unsupervised domain adaptation based on information-retrieval models for context aggregation. Following successful work last year on applying these methods to named-entity recognition, they are working on improving entity resolution. Using information-retrieval models addresses some difficulties in graph-based semi-supervised learning: natural-language problems often lack well-motivated similarity functions with small numbers of parameters.

UMass continued to develop large joint models of NLP with hidden structured variables. By constraining hidden variables to adhere to tree structure, and marginalizing out this hidden structure to optimize performance on the end task, reliance on jointly annotated data or pre-processing with trained parsers is reduced. They applied their approach to relation extraction, and ran more experiments on SRL. On some datasets, for both SRL and relation extraction, their approach outperformed systems that use a syntactic parser trained on annotated data. This is remarkable, because the hidden syntax model was trained without syntactic annotation at all. This work was presented at the EMNLP-CoNLL 2012 conference [122]. In analyzing the hidden syntactic structures that emerge from optimizing on the semantic end tasks, one interesting finding is that for the case of German, one of the datasets where the hidden syntax model outperformed the trained parser, the hidden model extracted deeper NP and PP structures when the parser returned flat constituents.

UMass worked on their Belief and Expectation Propagation-based approach to JI, and on applying their methods to joint NER, relation extraction, and co-reference. They improved their isolated NER, RE, and co-reference models even further toward state-of-the-art performance.

UMass worked on a sparse version of belief propagation that uses a combination of domain sparsity and message priorities to compute marginals in an anytime manner. When applied to a joint model of NER and relations, they achieved significant improvements in speed.

UMass explored applications of column generation, a method for solving very large-scale linear programs, to inference in graphical models with large state spaces. They also explored using a different technique for solving complicated linear programs, dual decomposition (DD). This has been successfully employed in recent years for various NLP applications of JI. Practitioners often solve the DD objective using subgradient descent, a method that is easy to implement and analyze, but may require many iterations to converge. An alternative method, block coordinate descent, has been used by others for DD-based message-passing schemes for inference in graphical models. In this work, it is assumed that maximizing the DD sub-problems is efficient. However, UMass targets NLP applications with structured sub-problems such as tagging and parsing, for which inference is computationally expensive. Therefore, the block-coordinate-descent scheme needs to be designed to minimize calls to the sub-problems. UMass devised such a scheme for tagging of sentences where the sentences are linked by global consistency constraints and preliminary experiments demonstrate a substantial speedup.

4.6.5 University of Wisconsin Madison and Wake Forest University

Profs. Shavlik, Re, and Natarajan. One of the major issues faced by Wisconsin while using the MR-KBP and TempEval 2010 datasets was the issue of *unannotated positive examples*. This is a prevalent issue as, for any human labeling, it may not be possible to get all valid event-time pairs annotated. As a result, during training, one cannot assume all the unannotated pairs to be negative examples. Hence, Wisconsin and Wake Forest collaboratively worked on extending their boosted models to handle missing labels. Inspired by the EM algorithm used in the literature to handle missing data, they developed an EM approach for learning the structure in relational models using functional gradient boosting. This extends their previous work on structure learning with completely observed data for two popular relational models: Relational Dependency Networks and MLNs. They derived the EM update equations along with approximations that make this approach feasible for relational models. They evaluated this approach on various relational datasets and showed that it is possible to learn effectively with missing data. This work is described in an ICML 2012 paper [116][9].

Wisconsin completed its study of its inference engine, Felix, which scales up MLN inference using a technique from mathematical programming called dual decomposition (DD). A standard approach for DD first partitions a graphical model into multiple tree-structured sub-problems. Wisconsin applied this approach to Markov Logic and showed that DD outperforms prior inference approaches on several tasks. Nevertheless, this standard approach is suboptimal, as it does not exploit the rich structure in the Markov Logic program. Thus, Wisconsin proposed a novel decomposition strategy that partitions a Markov Logic program into parts based on the structure of the program. A crucial advantage is that one can use specialized algorithms for portions of the input problem—some of which have been studied for decades e.g., co-reference resolution). Wisconsin performed extensive experiments to show that this program-level decomposition approach outperforms both non-decomposition and graphical model-based decomposition approaches to Markov Logic inference on several tasks. Felix is publicly available at http://hazy.cs.wisc.edu/hazy/felix/ and described in [85][10].

Wisconsin documented their experience in the TAC-KBP and MR-KBP challenges. Specifically, Wisconsin's scalable statistical inference system, Felix, allowed Wisconsin to rapidly integrate a diverse set of features and signals into their system. Thanks to this infrastructure, Wisconsin achieved state-of-the-art quality in TAC-KBP's test bed within several months through *Elementary*, a prototype knowledgebase-construction (KBC) system that is able to combine diverse resources and different KBC techniques via machine learning and statistical inference to construct knowledgebases.

- Using Elementary, Wisconsin implemented a solution to the TAC-KBP challenge with quality comparable to the state-of-the art, as well as an end-to-end online demonstration that automatically and continuously enriches Wikipedia with structured data by reading millions of webpages on a daily basis.

[9] http://ftp.cs.wisc.edu/machine-learning/shavlik-group/khot.srl12.pdf

[10] http://hazy.cs.wisc.edu/hazy/papers/felix-tr.pdf

- To take advantage of diverse data resources and proven techniques, Elementary employs Markov Logic, a succinct yet expressive language to specify probabilistic graphical models. Elementary accepts both domain-knowledge rules and classical machine-learning models such as conditional random fields, thereby integrating different data resources and KBC techniques in a principled manner.
- To support large-scale KBC with terabytes of data and millions of entities, Elementary leverages high-throughput parallel computing infrastructure such as Hadoop, Condor, and parallel databases. Further, to scale sophisticated statistical inference, Elementary employs a novel decomposition-based approach to Markov Logic inference that solves routine subtasks such as classification and co-reference with specialized algorithms.
- Elementary, through Felix, incorporates several novel and state-of-the-art techniques to perform very-large-scale inference, including dual decomposition, scoping rules and parameterized rule weights.
- Wisconsin empirically showed that this decomposition-based inference approach achieves higher performance than prior inference approaches. They conclusively demonstrated that its ability to incorporate diverse signals has positive impacts on KBC quality. This work is described in a journal article [145][11].

Wisconsin took advantage of their infrastructure of scalable joint inference to investigate a diverse set of problems in Natural Language Understanding (NLU). Wisconsin built a Wikipedia-based application, called DeepDive, to demonstrate their progress.

- DeepDive reads hundreds of millions of webpages, hundreds of thousands of web videos, books, and lectures to enrich Wikipedia. All deep NLU jobs (including named-entity recognition with the Stanford CoreNLP and dependency parsing) as well as statistical inference finished in several days using their scalable infrastructure. Wisconsin was able to set up a crawler that fetches millions of fresh news webpages on a daily basis, and use this data to continuously refresh their demonstration system DeepDive. At the time of writing, DeepDive contains three million entities, seven billion entity mentions, and one hundred million relations. DeepDive is available at http://research.cs.wisc.edu/hazy/deepdive/ with an overview at http://research.cs.wisc.edu/hazy/wisci.
- Wisconsin further applied their DeepDive experience to different application domains. They collaborated with geologists to setup a demonstration system called GeoDeepDive. GeoDeepDive extracts relationships between geologic formations (e.g., part of a mountain) and measurements (e.g., meters, percentage, etc.). After aggregating extractions, GeoDeepDive can estimate the total amount of carbon in different geologic areas of United States. GeoDeepDive uses the same backend processor as DeepDive, and is a promising indication that DeepDive's experience applies to other domains. A GeoDeepDive demo is available at http://hazy.cs.wisc.edu/hazy/geodeepdive/.

[11] http://ftp.cs.wisc.edu/machine-learning/shavlik-group/niu.ijswis12.pdf

Wisconsin performed the first study of how big data and information from crowdsourcing compare to each other at large scale. This study required processing the Web to understand the limits of both approaches. More specifically, Wisconsin performed a systematic study on how the sizes of two types of cheaply available resources impact the result quality of distant supervision, an increasingly popular technique for relation extraction: (1) unlabeled text corpora and (2) crowd-sourced human feedback (the crowdsourcing runs did not use DARPA funds).

- They found that text-corpus size has a stronger impact on the quality of distant supervision compared to human feedback. They also observed that distant supervision systems are often recall-limited due to the sheer variety and sparsity of natural language text that expresses a specific relation. Their results suggest that, to improve distant-supervision quality, one should first try to enlarge the training corpus, to increase recall, and then increase precision.
- They also observed that, when using the human labels alone to train relation extraction models, the "test set" quality is at approximately the same level as when using (a larger number of) noisy, distant-supervision labels. Thus, techniques that improve the quality of human-provided training examples are an interesting direction for future work.

Their ACL 2012 paper on this work provides more details [121][12].

Wisconsin investigated the task of maintaining the computations of sophisticated information-extraction techniques as new documents arrive. In particular, Wisconsin examined how to maintain conditional random fields (a de facto standard technique for many NLU subtasks including named-entity recognition and part-of-speech tagging). The lessons from this work are currently being utilized in the above-mentioned DeepDive demo. A paper describing this work appeared in ICDE 2012 [100][13].

Wisconsin developed a basic temporal annotation system for MR-KBP using the dependency paths between event and time expression. The dependency graph for a sentence was constructed using the Stanford NLP toolkit, where each edge has a dependency type and the dependency path is the path in this graph from the head word of the event to that of the time expression. Given the lack of training data in KBP, they decided to evaluate their system using TempEval 2010 dataset, which has similar temporal relations to the MR-KBP task, but with more labeled data. Evaluation of their system revealed that, in general, valid temporal relations have no verbs along the dependency path between the event and time expression. However, Wisconsin further discovered that for some special event-time pairs, verbs are found along the dependency path, but with specific dependency types such as *ccomp*, *partmod* attached to these verbs in the dependency path. They added a rule to allow verbs in the dependency path between the event-time pairs but let the learning algorithm discover the special dependency types that should be allowed [98].

[12] http://ftp.cs.wisc.edu/machine-learning/shavlik-group/zhang.acl12.pdf

[13] http://research.cs.wisc.edu/hazy/papers/crflex-icde12.pdf

Wisconsin developed *MDML*, a novel approach to performing metric learning via mirror descent. Metric learning is fundamentally concerned with how to compare two training examples, and a notion of similarity between them. Recently, metric learning methods have been applied extensively to large-scale text tasks such as text classification and document clustering. Most metric learning methods are characterized by diverse loss functions and projection methods, which naturally begs the question: is there a wider framework that can generalize many of these methods? In addition, ever-persistent issues are those of scalability to large datasets and the question of kernelizability.

Wisconsin developed a unified approach to metric learning: an online, regularized metric learning algorithm based on the ideas of composite objective mirror descent (COMID). The metric learning problem is formulated as a regularized, positive, semi-definite matrix-learning problem, whose update rules can be derived using COMID. This approach aims to be admissible to many different types of Bregman and loss functions, which allows for the tailoring of several different classes of algorithms. The most novel contribution is the use of the trace norm, which yields a sparse metric in its eigenspectrum, thus simultaneously performing feature selection along with metric learning. Wisconsin's initial empirical evaluation on benchmark datasets demonstrated that MDML learns comparably to existing approaches, but is several orders of magnitude faster. Details are described in a paper at ECML PKDD 2012 [147][14].

Wisconsin continued working on approaches to large-scale inference and optimization.

- They worked with Wisconsin's Condor group to develop a backend system that can distribute NLP jobs to thousands of machines on the National Open Science Grid. The outcome is a scalable batch-processing system that can harvest more than 100K machine hours in less than one week. This demonstrated the ability to deploy sophisticated rich models that are capable of improving the accuracy of their reading system, notably topic models and richer co-reference structures.
- Wisconsin is assisting the Knowledgebase Acceleration track in NIST's TREC 2012. By leveraging their scalable infrastructure, Wisconsin has been collaborating with the KBA (Knowledgebase Acceleration) team to process hundreds of millions of web documents with deep NLU tools such as Stanford NER.
- Wisconsin received donations due to their infrastructure. A KBA team with connections to the Applied Physics Laboratory at Johns Hopkins donated a dataset with hundreds of millions of documents to enhance DeepDive. Wisconsin also received a donation of video storage (NAS servers) from Johnson Controls, Inc.

[14] http://ftp.cs.wisc.edu/machine-learning/shavlik-group/kunapuli.ecml12.pdf

Wake Forest developed a preliminary version of an example creator for NL tasks to address a prevailing issue with NL problems: the paucity of labeled examples. They implemented an example creator that creates “low-weight examples” for Wisconsin's learning algorithms (RDN-Boost/MLN-Boost). At a high-level, this tool has three steps. In the first step, a query is constructed by the user (for example, the query “was born in”). The tool then uses the Lucene search engine to search the TAC-KBP corpus and retrieve the relevant documents containing this phrase. Second, a set of the most relevant documents is presented to the user, from which the user can select a sentence. This sentence chosen can be used as a template for retrieving further examples. In the final step, the tool constructs first-order logic predicates in the form requested by the user using the Stanford NLP toolkit to perform entity resolution and co-reference resolution in order to identify the examples.

Wake Forest collaborated with Wisconsin to develop a query-answering method based on the intuition that certain types of queries are easy to answer from the information extracted from Wikipedia. The approach combines information from Wikipedia infoboxes along with learned models (relational dependency networks, RDNs) to answer queries such as “When was Abraham Lincoln born?” In other cases, text from Wikipedia can be used to compute intermediate answers for answering more complex queries such as “Who succeeded the 35th President of United States?” In this case, Wikipedia can provide the name of the 35th President, and the learned model can use the “successor” concept to identify the next President’s name. Wake Forest and Wisconsin are working closely on tightly integrating the Boosting framework with the Infobox extractor that Wake Forest developed.

Wake Forest collaborated with SRI on implementing the ALBP algorithm. Standard lifted inference approaches try to avoid extensive propositionalization of first-order logic models through shattering (that is, decomposing the random variables into clusters of variables that exhibit identical behavior). In general, they require the entire model in order to compute a query's belief, and this requires the entire model to be shattered, which can be significantly expensive. ALBP differs in two key ways: first, the model is gradually shattered during inference so that only a portion of it is used for reasoning; and second, exact bounds on beliefs (the confidence interval) are derived. The latter is especially efficient and advantageous when only an approximate answer is needed, given that confidence intervals can be returned anytime during execution. The true marginal probability of the query will always be within this bound, and the limits of the bound converge tightly to this exact belief as shattering continues to include the entire model; longer execution times lead to tighter bounds. This property is essential when dealing with NL tasks because the evidence set is usually large and noisy.

Wisconsin studied efficient statistical inference for factor graphs that are larger than main memory. They implemented a prototype system that runs Gibbs sampling for factor graphs using different storage back-ends (e.g., raw files and HBase). They studied how classic database trade-offs can be adapted in the scenario of Gibbs sampling, and reported up to orders of magnitude speed-up by choosing the right plan in the trade-off space. This work was submitted to a top-tier database conference.

4.6.6 University of Illinois Urbana-Champaign (UIUC) (Prof. Roth)

4.6.5.1 Relation and Event Extraction for MR: IC++

UIUC developed a new model for jointly extracting argument roles of events from texts. UIUC's approach is designed to recognize and parse events in an unsupervised way, given only the events' definitions. The model takes events' definition in the form of event templates, along with coarse mention and type information for the event arguments. UIUC models the problem using a novel sequence-labeling model based on the latent-variable semi-Markov conditional random fields, addressing the event-extraction problem in a primarily unsupervised setting, where no labeled training instances are available. UIUC built on their work on Constraints Driven learning and proposed a learning framework called structured preference modeling that allows arbitrary preference to be assigned to certain structures during the learning procedure. Preference can be viewed as constraints that are in the form of properties, or templates of events.

Empirically, UIUC showed that this model, trained without annotated data and with a small number of structured preferences, yields performance competitive to some baseline supervised approaches. This work appeared in ACL-12 [120].

4.6.5.1.1 Joint Inference and Learning: Constrained Conditional Models (CCMs)

UIUC published a paper [109] describing their Phase 3 development of a general framework containing a graded spectrum of Expectation Maximization (EM) algorithms called Unified EM. UIUC further developed a framework for learning an inference while controlling the entropy of the distribution over predicted output, to be used as a measure of confidence in the structured output prediction. This paper appeared at an ICML-12 workshop [115].

UIUC proposed a new model for decomposed structured learning. Unfortunately, in structured prediction settings with expressive inter-variable interactions, exact inference-based learning is often intractable. UIUC developed a new way, Decomposed Learning (DecL), for performing efficient learning over structured output spaces. The key idea is that in DecL, one restricts the inference step to a limited part of the output space. UIUC used characterizations based on the structure, target parameters, and gold labels to guarantee that DecL with limited inference is equivalent to exact learning. UIUC showed that in real-world NLP settings, DecL-based algorithms are significantly more efficient and provide accuracies close to exact learning. This paper appeared in ICML-12 [113].

UIUC made significant progress developing an amortized ILP algorithm. Typically, in structure prediction, an inference procedure is applied to each example independently of the others. UIUC tried to optimize the inference time complexity over entire datasets, rather than individual examples. UIUC proposed three exact inference theorems that enable reusing earlier solutions for certain examples, thereby completely avoiding possibly expensive calls to an ILP solver. UIUC also identified several approximation schemes that can provide further speedup. UIUC instantiated these ideas to the structure-prediction task of semantic role labeling and showed that one can achieve a speedup of more than 2.5 times using this approach while retaining the guarantees of exactness and a further speedup of over 3 times using an approximation that does not degrade performance. This work appeared in EMNLP/CoNLL-12 [132].

4.6.5.1.2 Extended Semantic Role Labeling

A key component in UIUC's EE approach is an extended semantic role-labeling methodology, providing generic semantic parsing: verbal, nominal, and prepositional predicates are identified and their arguments are assigned their roles with respect to the predicates.

UIUC worked on integrating nominal relations, verb-based relations, and prepositional-based relations. UIUC investigated several global-inference approaches to support this process.

UIUC had a semi-breakthrough in this direction, based on an improved latent learning approach, and given some refinement of the type of argument roles that were developed to provide roles that are more coherent. UIUC can now show that this joint inference approach yields significant improvement in preposition role identification. In addition to the learning breakthrough, UIUC has developed a taxonomy for preposition-based relations. The work describing UIUC's progress on the SRL for preposition was accepted to TACL but won't appear until after project end.

4.6.5.1.3 Reference (Grounding) of Concepts and Entities (Wikification) and Co-Reference Resolution

UIUC participated in the CoNLL-12 Shared Task on co-reference resolution and their submission was one of the top few English submissions. The key innovation in UIUC's submission is a new learning algorithm for co-reference resolution; while the problem is a structured prediction problem, standard structured prediction algorithms like Structured SVM cannot scale enough for large documents. UIUC developed a stochastic version of Structured SVM that is as fast as standard classification algorithms. A short version of this contribution appeared at EMNLP/CoNLL-12 Shared Task Proceedings [133] and a longer version is in preparation. UIUC developed a better understanding of the learning processes involved in co-reference resolution and, in particular, of the role of latent representations in the learning. This understanding is shown to yield good improvements in the performance on learning based co-reference resolution.

4.6.5.1.4 Event Temporal, Causal Reasoning, and Timelining

Event Timeline: UIUC developed a JI algorithm for constructing a timeline of events mentioned in a given text. To accomplish that, UIUC suggested a new representation of the temporal structure of a news article based on time intervals. UIUC then presented an algorithmic approach that jointly optimizes the temporal structure by coupling two local models: (1) a model that predicts associations between two events and (2) a model that maps events to the temporal interval they occurred in. The global inference makes use of global constraints over events, relations between them and temporal intervals. Moreover, UIUC presented ways to leverage knowledge provided by event co-reference to further improve the system performance. Overall, experiments show that this JI model significantly outperformed the local model, and that the use of good event co-reference could make a remarkable contribution to a robust event timeline construction system. This work appeared in EMNLP/CoNLL-12 [132].

Temporal Reasoning: The work mentioned above builds on event-identification capabilities and on augmented extraction capabilities of temporal intervals. UIUC supports extracting temporal phrases, normalizing them to a canonical representation and recognizing basic relations between temporal intervals; UIUC currently use six types of relations (before, after, overlaps, etc.). UIUC's system was presented as a demo paper at NAACL-12 [108]. UIUC gave a tutorial on temporal extraction and reasoning at COLING'12.

4.6.5.2 NLP Tools and Software Architecture

SRI and UIUC updated the Curator install and continued testing and debugging efforts. This included creating a new custom Curator annotator to explore issues regarding how parallel annotators are utilized and situations where processing is not correctly distributed. This work was presented at LREC-12 [103]. A new version was announced to the research community.

UIUC improved the Event Annotation Tool (EAT+) given feedback from SRI. The tool is being used now for annotation of events and temporal information. UIUC evaluated the event extraction output from Curator against reference annotations produced by humans using the EAT+ tool.

4.6.5.3 Textual Inference

UIUC is interested in inference over information stated in human language, which they characterize in terms of comparing spans of text (phrases, sentences, paragraphs) and determining whether they express the same information, and if different, in what way they are different (for example, whether they contradict or complement each other). The goal is to determine whether one statement holds in, contradicts, complements, or is completely unrelated to another. UIUC focused this work on textual inference with respect to complex verb constructions.

UIUC developed a lexical textual entailment (TE) system with a light-verb constructions (LVCs) identifier and investigated the effectiveness of detecting LVCs within this TE system. UIUC generated and annotated a TE corpus specifically tailored for English LVCs, and showed that the ability to classify LVCs in a given context contributes to 39.5% error deduction in accuracy and 21.6% absolute F1 value improvement in supporting this type of inference without attending to LVC in this sophisticated way. UIUC also considered the identification of phrasal verbs and making use of it in the context of a textual entailment framework. In particular, UIUC investigated a supervised machine-learning framework for automatically identifying English phrasal verbs in a given context. UIUC concentrated on those phrasal verbs that are defined as the most confusing phrasal verbs, in the sense that they are the combinations of the most common verbs, such as “get,” “make,” and “take,” and the most frequent prepositions and particles, such as “up,” “in,” and “on.” This work appeared in SEM-12 [110].

4.6.7 UIUC (Prof. Amir)

Prof. Amir’s team worked on probabilistic modal (PM) operators for natural language understanding. They investigated using PM models to represent what authors assume about readers knowledge, and created a theoretical framework for inferring Bayesian-Network PM models from text, and an implementation of that framework in computer algorithms and executable programs.

UIUC extended Probabilistic Modal models to dynamic domains in which actions change the state of the world. These models capture events in natural-language texts, and enable modeling the beliefs of authors about beliefs of readers about those events and their participants. UIUC used an action-based dynamic model where effects of actions are modeled as stochastic choice between deterministic executions.

UIUC built a specification language that represents changes and observations in a probabilistic world, and implemented a game-theory-based engine that reasons about an agent’s beliefs about other agents’ probabilistic beliefs. This work appeared in AAAI’12 [134, 135], and in [144].

4.6.8 University of Washington (UW)

Profs. Domingos and Zettlemoyer. UW developed TML, a tractable subset of Markov Logic that can be used for logical-probabilistic representation and tractable joint inference over the entire machine-reading process, including syntactic and semantic parsing, ontology and knowledgebase population, and question answering. A paper on this appeared in AAAI-12 [156][15] .

UW worked on an algorithm for coarse-to-fine variational inference using sum-product networks. This will be the basis of an algorithm for approximate inference when the knowledge acquired translates into an overly large knowledgebase in TML. In other words, when the KB is too large, UW variationally finds the closest tractable one and uses it instead.

UW developed an algorithm for multiple hierarchical relational clustering. This will be used to induce and populate a consistent TML ontology from the raw unresolved facts extracted from the text. Ongoing work beginning in this phase is focused on adding the ability to reason robustly about background knowledge to provide partial supervision for the induced ontology, to support open semantic parsing approach.

UW continued development of their integrated development environment for the rapid debugging of rule-based extractors with learning over interactively defined features. A new interface and optimized execution routines allow for near instantaneous extractions over very large datasets, greatly improving usability. UW completed an initial version and began working on evaluation. UW found that highly accurate extractors can be built by an expert user in under an hour, for each of the initial five relations considered. Ongoing work is focused on expanding this evaluation and writing up the results.

UW began work on defining new models for scalable, open semantic parsing. UW worked on building a dataset for scalable question answering against Internet-scale databases, such as Freebase and DBpedia. The questions include sentences from the recent QALD-2 dataset and a subset of questions from the TREC fact and list competitions. UW also began development of initial models for learning database-independent semantic parsers, which could be trained with minimal supervision for any specific database.

UW continued ongoing work on building a linear-time shift-reduce CCG semantic parser. UW developed a new framework and heuristics for A* parsing that has the potential to, for the first time, provide provable correct results with linear time performance in sentence length. Implementation is ongoing.

[15]. http://homes.cs.washington.edu/~pedrod/papers/aaai12.pdf

4.6.9 Onyx Consulting

Prof. Nirenburg. Ellipsis is a linguistic process that renders certain aspects of text meaning invisible at surface structure, thereby making them inaccessible to most current text-processing methods. Ellipsis is considered one of the more difficult aspects of text processing and, accordingly, has not been widely pursued in NLP applications. However, not all cases of ellipsis are created equal: some can be detected and resolved with high confidence within the current state of the art. Onyx has been working toward configuring a system that can resolve one class of elliptical phenomena: *elided scopes of modality*. Onyx has addressed the problem of elided scopes of modality from two perspectives:

(1) Onyx developed a full microtheory of modal-scope ellipsis treatment that will be incorporated into the language-enabled intelligent agents in the OntoAgent cognitive architecture. This work is reported in the conference paper “Resolving Elided Scopes of Modality in OntoAgent” [151], which was presented at the First Annual Conference on Advances in Cognitive Systems (http://www.cogsys.org/). This approach employs all of the static knowledge resources and reasoning engines available to OntoAgent intelligent agents.

(2) Onyx developed a method for detecting and resolving a subset of cases of modal-scope ellipsis that can be applied to big data. To work over big data in real time, the approach uses only a subset of the resources and reasoners available in this environment and replaces some of the more resource-intensive aspects of processing with cheaper proxies. The goal was to focus on achieving high precision over a large dataset.

As shown in Figure 10, the system takes as input Onyx’s indexed version of the Gigaword corpus and selects examples that include modal scope ellipsis. Those examples are analyzed by a

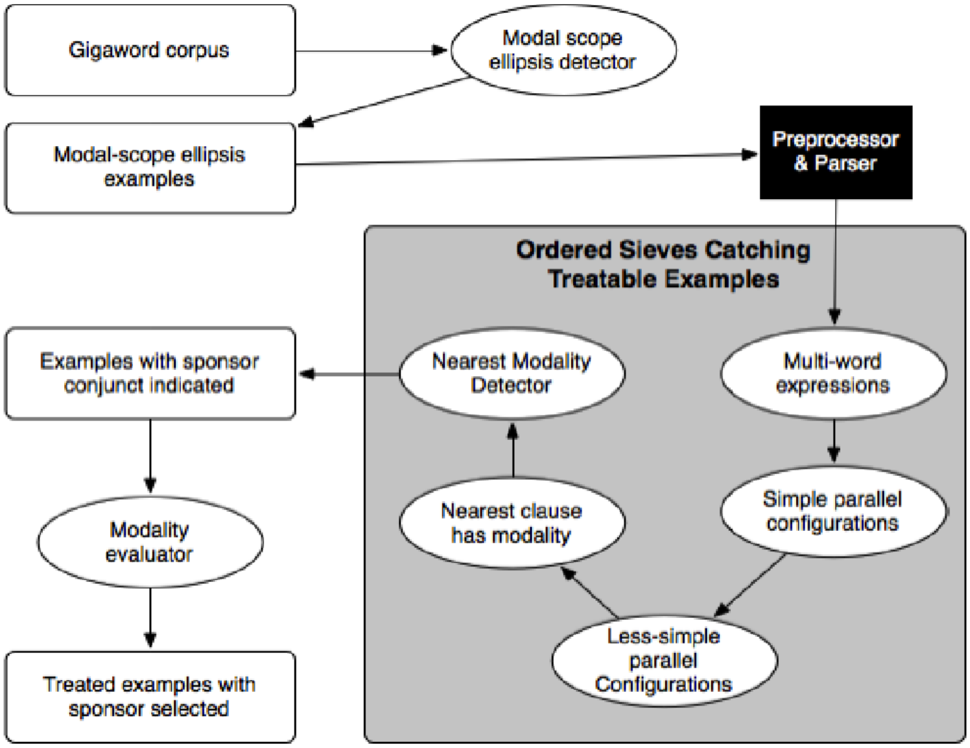

Gigaword corpus
Modal scope ellipsis detector
Modal-scope ellipsis examples
Preprocessor & Parser
Ordered Sieves Catching Treatable Examples
Examples with sponsor conjunct indicated
Nearest Modality Detector
Multi-word expressions
Nearest clause has modality
Simple parallel configurations
Modality evaluator
Less-simple parallel Configurations
Treated examples with sponsor selected

Figure 10: Detection and Resolution of Modal Scope Ellipsis

preprocessor and parser that, for purposes of this experiment, are treated as black boxes. The next series of engines, which use heuristic evidence from preprocessing and parsing, act as sieves (cf., e.g., [91] Ratinov & Roth, 2012, for the sieve metaphor), each one catching examples of a particular profile to treat. The output of the sieves is a pointer to the text span that is believed to contain the sponsor. Once the system knows where to look for the sponsor, it needs to evaluate whether any modalities contained therein should be included in, or excluded from, the sponsor. This work is carried out by the Modality Evaluator. The output of this engine is a set of examples decorated with metadata indicating how to resolve the elided scope of modality.

To summarize, Onyx Consulting developed a "mini-microtheory" of modal-scope ellipsis resolution that could be applied to big data. The approach was developed iteratively using evidence from the Gigaword corpus. Details about the theory and implementation are provided in Appendix D, written by Onyx Consulting, Inc.

4.6.10 Columbia University

Prof. Collins. Columbia University further developed the spectral learning algorithm for latent-variable PCFGs (L-PCFGs), in particular implementing experiments with this method. A paper on these experiments has been accepted for publication at NAACL 2013. The experiments show that the method performs at the same accuracy as EM, but is around 20 times faster to train (roughly speaking, the method has the same cost as a single iteration of EM; EM takes around 20–30 iterations to converge to a good solution). IHMC describes a number of key steps in obtaining this level of performance. A simple backed-off smoothing method is used to estimate the large number of parameters in the model. The spectral algorithm requires functions mapping inside and outside trees to feature vectors—making use of features corresponding to single-level rules, and larger tree fragments composed of two or three levels of rules. IHMC shows that it is important to scale features by their inverse variance, in a manner that is closely related to methods used in canonical correlation analysis. Negative values can cause issues in spectral algorithms, but a solution is described to these problems.

4.6.11 Institute for Human and Machine Cognition (IHMC)

Prof. Wilks. IHMC executed an exploratory project to locate proto-beliefs of individual Ummah message board posters on a large scale. These beliefs could then be examined to determine the consistency of an individual poster's beliefs and to identify where that individual's beliefs conflict with the beliefs of others; such conflicts of belief could occur either within the context of a single thread or in the context of all threads.

In the information flow of the completed system, facts were extracted from the Ummah message board postings using unsupervised methods for information extraction. These facts were then linked to individual posters as beliefs or assertions in a belief management engine. Finally, heuristics were used to investigate confirmations and negations of beliefs within and outside individual message threads.

As an exploratory effort, the project's aim was to determine the feasibility of IHMC's approach to extraction and comprehension of agents' interrelated beliefs. The primary outcome of the completed work is a positive demonstration of the extraction of these beliefs. In particular, IHMC demonstrated that (1) beliefs could be extracted from the unstructured data contained in an online forum, (2) represented in the ViewGen belief engine, and (3) scored using heuristic approaches similar to the FactRank (Jain & Pantel, 2010) algorithm.

IHMC developed several experimental algorithms for integrating the FactRank fact-confirmation algorithm into ViewGen's core ascription algorithm. IHMC's initial "best subset" algorithm, which ascribes the highest scoring subset of consistent beliefs, performs well when beliefs are sparse but is not tractable when the belief space is dense. IHMC integrated FactRank as a scoring metric in their "greedy" ascription algorithms. While IHMC learned a lot about the use of random-walk scoring algorithms, several questions remain that are of importance to the use of such algorithms in a "belief confirmation" context:

(1) How does one properly score contradictory facts (say, P and not-P) versus the simple falsity (or non-confirmation) of a fact (say, that P is not true or not confirmed)?

(2) In the context of beliefs and differing viewpoints, can beliefs be scored en masse regardless of viewpoint or should the beliefs of one agent be scored independently and in isolation from the beliefs of other agents?

(3) Given that the score of individual facts/beliefs is based on the overall graph, how brittle are rankings to topological changes—specifically, how do the ordered rankings of facts within an arbitrary sub-graph compare to a rescoring of that sub-graph as its own independent graph?

With the successful conclusion of this exploratory project, the research effort is being extended and expanded as part the DARPA DEFT project, where IHMC will address the above questions. In Appendix D, IHMC provides motivating background and then details the technical tasks and activities comprising the conducted research.

To summarize, IHMC developed a prototype system capable of extracting, modeling, and scoring beliefs that are assigned to forum posters and for representing posters' reflexive beliefs of themselves and others.

4.6.12 SRI International

4.6.12.1 SRI Research Team

Dr. Rodrigo de Salvo Braz. SRI's research team developed a powerful probabilistic inference engine and was in the process of applying it to joint inference for NLP as the project ended. In particular, this engine was applied to the joint problem of named-entity recognition, co-reference, and relational extraction from text by writing models that contained both linguistic and domain knowledge.

Relational (or first-order) Probabilistic Representations: With this type of representation, the model is specified in a compact manner, with rules that hold for multiple random variables in the domain. For example, instead of specifying a conditional probability from *movedTo*(john, kitchen) and *isAt*(john,kitchen), another identical one for *movedTo*(mary, school) and *isAt*(mary,school), and so on, our representation enables a more compact and natural generic rule on *movedTo*(Person, Place) and *isAt*(Person, Place), which is represented only once.

Lifted Probabilistic Inference: Using relational probabilistic representation helps with compact model specifications, but standard inference techniques require that representation to be instantiated into regular (propositional) conditional probabilities (or potential functions, in the case of undirected models), losing compactness and generating a large model. *Lifted probabilistic inference*, on the other hand, manipulates the representation in first-order form, keeping it compact and performing operations on a single conditional probability function. By contrast, regular inference would perform the same operations repeatedly, for each instance of that function, thus producing exponential effort.

Anytime Lifted Probabilistic Inference: Exact (and many approximate) inference algorithms need to examine an entire model before making predictions, even if lifted. However, it is often the case that a query's answer mostly depends on a very small fraction of a model. When a model represents a very large collection of knowledge, examining its entirety when only a small fraction is fundamentally necessary is very wasteful. *Anytime lifted probabilistic inference* is an incremental inference method that updates a query's answer gradually as it examines increasingly relevant portions of the model. If the query depends on only a small fraction of the model, as most do, then the algorithm will not need to examine the entire model to answer.

Model Counting of Equality Formula Constraints: Lifted-inference algorithms represent generic constructs during their operation that stand for an entire set of probabilistic concepts. For example, they may need to represent the set of conditional probabilities *P(isAt(Person, Place) | movedTo(Person, Place)) for Person ≠ john* (perhaps because we have specific knowledge about John, for example). Because a form of unification takes place during the algorithm's operation, equality constraints of this sort need to be manipulated, and, in fact, one must keep track of how many *solutions* they have. This is an important sub-problem of lifted inference for which no satisfactory solution had been offered until recently, so we developed a solution for it, described in a paper at a UAI-12 workshop [139].

Probabilistic Inference as Symbolic Evaluation: Perhaps because lifted inference is a relatively new area, and because it involves a more abstract level of description, algorithm descriptions in the literature have often been confusing and ambiguous. SRI developed a formal notation and representation to describe them without ambiguity. The benefits go further, because it enables casting lifted inference as a form of *symbolic evaluation*, that is, an evaluation of mathematical expressions in which not all variables have known values. This is useful, for example, in performing computations with a conditional probability *P(isAt(Person, Place) | movedTo(Person, Place)), Person ≠ john*, even though the variables *Person* and *Place* do not have a specific value assigned to them, but only *constraints* on their values (such as *Person ≠ john*). Explicitly representing and manipulating the mathematical expressions composing the problem opens up possibilities for describing anytime lifted inference in a simpler manner, as it can be viewed as a sort of lazy evaluation of these expressions.

SRI implemented a Lifted Belief Propagation (LBP) algorithm as symbolic evaluation. Unlike other versions of LBP, this version takes the internal structures of factors into account, because they are represented as regular mathematical expressions. The anytime version currently only works with the non-loopy version of the algorithm.

SRI has released the software of the probabilistic inference engine, the symbolic evaluation system, and general utilities, as three separate projects. The sites contain the code, detailed documentation, and a wiki, and can be found at:

https://code.google.com/p/aic-praise/

https://code.google.com/p/aic-expresso/

https://code.google.com/p/aic-util/

SRI worked on inference in the presence of symmetry. The presence of non-symmetric evidence has been a barrier for the application of lifted inference (an inference method that first identifies sets of objects that are symmetrical given the model, then performs computations on those symmetrical sets of objects instead of individual objects) because the evidence destroys the symmetry of the model. In the extreme case, if distinct facts are observed on each individual in a group then all current lifted inference methods reduce to traditional ground inference methods whose complexities are exponential in the number of individuals.

SRI developed a new lifted inference method, LIDE (Lifted Inference with Distinct Evidence), that allows polynomial-time exact lifted inference even in the presence of unique evidence on a set of grounding instances of a unary predicate, one for each individual. Instead of shattering (that is, breaking symmetries in) the original model with the evidence, as previous lifted inference methods do, LIDE applies lifted inference to the unshattered model to obtain the marginal probability on the sets of symmetrical random variables on which we have evidence. Because the model is unshattered, this calculation is polynomial on the size of these sets. Then this marginal probability and the evidence are used to compute a posterior probability as well as the maximum a posteriori (MAP) configuration in polynomial-time.

Experiments on the “Friends & Smokers” MLN show that LIDE can perform exact inference much faster than (lifted) belief propagation (BP), which is a very efficient approximate method, without suffering from non-convergence issues or approximation errors. For example, in the case of 800 people, LIDE took 132 seconds while BP took 643.2 seconds (the time for grounding the network is not included in the BP running times). As a result, within 15 minutes, BP can run only up to the case of 800 people while LIDE can run up to the case of 1500 people. This work is described in a AAAI 2012 paper [138][16].

SRI, in collaboration with UMass, developed a general framework for lifting variational approximation algorithms such as LP relaxation of MAP inference, a widely used approximation in NLP problems. Our new approach, based rigorously on the theory of group action, introduces the concept of the *automorphism group* of an exponential family or a graphical model, thus provides the first formalization of the general notion of symmetry of a probabilistic model. This automorphism group provides a precise mathematical framework for lifted inference in the general exponential family. Its group action partitions the set of random variables and feature functions into equivalent classes (called orbits) having identical marginals and expectations. Then the inference problem is effectively reduced to that of computing marginals or expectations for each class, thus avoiding the need to deal with each individual variable or feature.

We demonstrated the usefulness of this general framework in lifting two classes of variational approximation for MAP inference: local LP relaxation and local LP relaxation with cycle constraints; the latter yields the first lifted inference that operates on a bound tighter than local constraints. Initial experimental results demonstrate that lifted MAP inference with cycle constraints achieved state-of-the-art performance, obtained much better objective function values than local approximation while remaining relatively efficient (order-of-magnitude faster than inference on the ground model). This work is described in a 2012 UAI paper [140].

4.6.12.2 SRI Software Engineering Team

Per DARPA’s guidance, the FAUST SE team stopped work on a single integrated system or more specifically, toward creating an end-to-end evaluation system, but continued efforts to assist in the development and maturation of research being conducted across the FAUST team though primarily in support of the SRI research team.

In Phase 3R, the FAUST SE team:

(1) Provided software-engineering advice and suggestions for enhancements to subcontractors. This included parallel-processing and caching enhancements in existing NLP tools. These tools can then effectively process larger amounts of data in smaller timeframes;
(2) Supported efforts for reasoning that make use of geographic information contained in SRI’s Gazetteer module, which we released as a general-purpose tool;
(3) Supported SRI’s Anytime Lifted Probabilistic Belief (ALPB) development. Primarily, we developed (1) a graphical user interface, (2) evaluation and experiment framework, and (3) associated web services. We provided an interactive environment for users, researchers running experiments, and developers of the ALPB system.

[16] http://www.ai.sri.com/~huynh/papers/bui_huynh_braz_aaai2012.pdf

(4) Implemented ALBP utilities, including ANTRL-based parsers for grammars, output converters to human-readable displays, unit tests, and scalability improvements.

5. CONCLUSIONS

MRP lasted for only three of the planned five phases, so we did not get to fully test our overarching hypothesis that a machine reading system based on joint inference over relational (first-order) models could be made tractable. However, we did make considerable progress towards that goal, and results so far are, at the very least, promising.

One key method for achieving our goals was selecting the best research team to realize our distinctive research vision. To this end, we assembled a team that included leading researchers in NLP, probabilistic reasoning, and machine learning (ML). Our progress toward our vision is made evident by our team's extensive contributions to scientific knowledge. FAUST researchers have won five best paper awards and have published 155 papers (so far), most in top conferences, for their papers on MRP-sponsored work (see Appendix A). In addition, the FAUST team developed extensive MR-related software that is freely available (see Appendix B).

A second key for achieving our goals was fully supporting the software integration tasks and MRP evaluations. This support included both an excellent and experienced Software Engineering (SE) team and sufficient funding for the integration and evaluation tasks, which are often underestimated. Helping the Government define the evaluations, preparing for them, and participating in them consumed a significant amount of our effort.

Our team made major advances and explored new directions in NL understanding, at levels ranging from providing general infrastructure components useful to many groups to cutting-edge research into new models of language. At the spectrum's practical end, a key result of MRP was the development of Stanford's CoreNLP, a simple-but-flexible pipeline framework that ties together all of Stanford's core NLP components, from sentence splitting and tokenization through parts-of-speech, named entities, to parsing and co-reference, and makes them available under a simple uniform API. CoreNLP was made publicly available open source.

Stanford developed a new deterministic sieve architecture for entity co-reference. This system was the best performing system at the CoNLL 2011 Shared Task [68] on entity co-reference. Stanford developed improved relation-extraction systems (finding semantic predicates and their arguments). This was explored using both fully supervised methods over linguistic analyses, as in the Phase 2 evaluation, and more extensively by considering the task of distantly supervised learning, where you have some texts and an initial knowledgebase that you wish to extend with more texts. Stanford worked on this problem extensively in the context of the NIST TAC KBP task, and developed a new, principled model that handles the uncertainties of distantly supervised learning, the Multi-Instance, Multi-Label Relation Extraction (MIML-RE) model.

Pushing the frontiers of research, Stanford concentrated in three main areas. First, they explored the usage of joint learning methods within NLP, doing things such as showing gains from doing joint named entity recognition and parsing, or doing successful joint learning over texts from different domains and genres. Second, they extended work on co-reference and event extraction, introducing a new model of cross-document joint entity and event co-reference. Finally, Stanford initiated a major exploration of deep learning (multi-layer neural network) methods for use of the data-dependent recursive hierarchical structures of natural language. This lead to the development of several new models for handling composition within vector spaces, NLP applications to parsing, sentiment analysis and relation classification, and the application of these methods to both vision and language, which won an ICML 2011 best paper award [70].

The University of Wisconsin developed modules for very-large-scale joint inference—Tuffy (a MLN RDBMS-based inference engine, which has been downloaded more than 5000 times) and Felix (an operator-based relational optimizer for statistical inference). They developed new approaches for very-large-scale inference, including optimization approaches such as dual decomposition and partitioning-based inference algorithms. Wake Forest collaborated with SRI to develop the Anytime Lifted Belief Propagation (ALBP) algorithm. Wisconsin demonstrated that their approach of using probabilistic logic to extract information from text scaled to a corpus of more than one billion documents.

Wisconsin and Wake Forest collaborated to develop novel approaches and algorithms to address several open problems in Statistical Relational Learning (SRL). These approaches were quite effective when applied to MR and other text-based datasets.

The University of Washington pioneered and extended a diverse set of approaches for distant supervision of relational extractors. Their methods used background knowledgebases ranging from Wikipedia, Freebase, and the Nell KB, and matched to a variety of textual corpora including Wikipedia and newswire text. Their LUCHS system generated extractors for more than 5000 distinct relations [21], which is several orders of magnitude more than previous systems. Their MultiR system included a novel graphical model that not only relaxes the common prior assumptions of disjoint relation tuples, but requires two orders of magnitude less computational time than previous multi-instance methods. Finally, their VELVET system introduced the notion of ontological smoothing, a method for quickly training a relational extractor with only a handful of positive examples.

UW (Prof. Domingos) also worked toward an end-to-end solution to machine reading that builds on top of unsupervised semantic parsing and enables large-scale JI that will be efficient enough to support the FAUST vision. UW's goals were three-fold: (1) to create new architectures and algorithms for efficient, large-scale probabilistic JI; (2) to develop algorithms that unify probabilistic and logical inference; and (3) to develop methods for scalable semantic parsing from text. They developed (1) USP, an algorithm for unsupervised semantic parsing, taking steps toward making it online and more scalable [1]; (2) the CFPI framework for coarse-to-fine probabilistic inference; (3) PTP, a new approach for unifying logical and probabilistic inference; (4) ABQ, a new approach for efficiently conducting approximate probabilistic inference; (5) SPNs, a new deep architecture that is more general than arithmetic circuits and also enables efficient exact inference; (6) a theory of USPN, an end-to-end solution to machine reading that would extend USP to process text online; (7) a family of deterministic, structured message-passing algorithms for efficient JI; (8) an algorithm for multiple hierarchical relational clustering; (9) TML, a tractable subset of Markov Logic that can be used for logical-probabilistic representation and tractable JI over the entire machine-reading process, including syntactic and semantic parsing, ontology and knowledgebase population, and question answering; and (10) a linear-time shift-reduce CCG semantic parser.

The University of Massachusetts Amherst developed a joint model for event extraction that combined entity-type prediction and detection of event arguments. This model ranked first in the BioNLP shared task. UMass built the first cross-document joint NER and relation-extraction model, trained only with weak supervision. UMass developed the FACTORIE toolkit for deployable probabilistic modeling and released their BioNLP event extraction.

UMass developed SampleRank, a highly scalable algorithm for learning in large-scale graphical models. This algorithm supports arbitrary, user-specified loss functions, and trains models both more quickly and more accurately than previous methods. UMass pioneered a new paradigm for distant supervision by introducing latent variables that indicate whether a relation is expressed by a mention. This new type of model has improved the accuracy of relation extraction.

UMass developed several JI algorithms to make JI scalable. These algorithms were orders of magnitude faster than previous methods. The speed was achieved by lazily instantiating both factors and variable values only when they were needed. UMass developed a new generation of cross-document co-reference algorithms that rely on hierarchies of co-reference clusters for both increased robustness and efficient parallel inference.

UIUC pioneered an ILP-based framework to support incorporating declarative knowledge as a way to guide learning and support global inference. They developed new algorithms for learning with indirect supervision, and for learning and inference with latent representations. The UIUC framework was used in developing multiple NLP capabilities, including (1) relation and event extraction; (2) co-reference; (3) textual inference; and (4) temporal and causal reasoning. Their key contribution to learning was the development of the Wikifier, an approach for disambiguating concepts and entities appearing in text and grounding them in an encyclopedic resource. This is both a knowledge-acquisition tool and a way to support co-reference within and across documents and other textual inferences. UIUC tools are available to the research community, including the Curator, a distributed system for running and aligning multiple-state NLP preprocessing tools, as well as state-of-the-art tools for multiple NLP tasks, including semantic role labeling, named entity recognition, and co-reference resolution.

UIUC (Prof. Amir) worked on probabilistic modal (PM) operators for natural language understanding. They investigated using Probabilistic Modal models to represent what authors assume about readers’ knowledge, and created both a theoretical framework for inferring Bayesian Network PMMs from text and an implementation of that framework in computer algorithms and executable programs.

PARC and CSLI's work focused on inferences that can be drawn from texts based on inferential properties of linguistic expressions. Such inferences are a necessary part of automated NL understanding. This work demonstrated that the task can be aided by different kinds of resources, including lexical class markings, ontological classifications, and domain models that link different classes of items together.

PARC/CSLI based their study of the veridicality inferences of texts on the following broad hypothesis: (1) a large class of lexical items in particular syntactic frames, or specific types of phrases, are associated with a veridicality signature; (2) the implications of whole sentences about their author's commitments arise from a projection mechanism from the veridicality signatures of the elements embedded in them; and (3) contextual factors might strengthen these implications. They developed and analysis of the range of veridicality signatures and the environments in which lexical items have them and verified the analysis with human subjects. They developed an algorithm of projection that took into account the effect of contextual factors.

MIT/Columbia developed novel methods based on dual decomposition and Lagrangian relaxation for inference in NLP. The method was shown to be effective in a number of NLP problems. The work resulted in several publications (including a best paper award at EMNLP 2010) [36]. They also developed novel spectral-learning methods for latent-variable models, and completed experiments showing that the methods perform at the same level of accuracy as Expectation Maximization (EM), which is widely applied in NLP, but are an order of magnitude more efficient in training time.

IHMC executed an exploratory project in Phase 3R to locate proto-beliefs of individual Ummah message board posters on a large scale. Facts were extracted from the Ummah message board postings using unsupervised methods for information extraction. These facts were then linked to individual posters as beliefs or assertions in a belief management engine. Finally, heuristics were used to investigate confirmations and negations of beliefs within and outside individual message threads. The primary outcome of the completed work is a positive demonstration of the extraction of these beliefs.

Ellipsis is a linguistic process that renders certain aspects of text meaning invisible at the surface structure, thereby making them inaccessible to most current text-processing methods. Ellipsis is considered one of the more difficult aspects of text processing and, accordingly, has not been widely pursued in NLP applications. Onyx worked in Phase 3R toward a system that can resolve one class of elliptical phenomena: elided scopes of modality. They developed a full microtheory of modal-scope ellipsis treatment [151] and a method of detecting and resolving a subset of cases of modal scope ellipsis that can be applied to big data.

SRI's research team developed a powerful probabilistic inference engine and used it on JI for NLP. Their *Lifted probabilistic inference* manipulates the representation in first-order form, keeping it compact and performing operations on a single conditional probability function where regular inference would perform the same operations repeatedly, for each instance of that function [138,139]. SRI developed an engine for *anytime lifted probabilistic inference,* an incremental inference method that updates a query's answer gradually as it examines increasingly relevant portions of the model. If the query depends only on a small fraction of the model, as most do, then the algorithm will not need to examine the entire model to answer [141].

SRI developed a new lifted inference method, LIDE (Lifted Inference with Distinct Evidence), that allows polynomial-time exact lifted inference. LIDE applies lifted inference to the unshattered model to obtain the marginal probability on the sets of symmetrical random variables on which we have evidence. Because the model is unshattered, this calculation is polynomial on the size of these sets [138].

6. REFERENCES

Numbered cites refer to our table of publications in Appendix A. These references are non-FAUST papers that are cited in long form, e.g., "(Rush and Collins, 2011)".

(Bobrow and Winograd 1977). An overview of KRL, a Knowledge Representation Language. Cognitive Science, Volume 1, Issue 1, January 1977, Pages 3–46.

(Chang et al., 2007). Chang, M., Ratinov, L., and Roth, D. Guiding semi- supervision with constraint-driven learning. In Proc. of the Annual Meeting of the ACL, 2007.

(Ferrucci et al. 2010), David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. Building Watson: An Overview of the DeepQA Project. AI Magazine, Fall 2010, pp. 59-79.

(Ganchev et al., 2010). Kuzman Ganchev , João Graça , Jennifer Gillenwater , Ben Taskar , and Michael Collins. Posterior Regularization for Structured Latent Variable Models, Journal of Machine Learning Research, Volume 11, pp 2001−2049, 2010.

(Hobbs, et al. 1993). Jerry R. Hobbs, , Mark E. Stickel, Douglas E. Appelt, Paul Martin. Interpretation as Abduction, in Artificial Intelligence, Volume 63, Issues 1–2, October 1993, Pages 69–142. http://dx.doi.org/10.1016/0004-3702(93)90015-4

(Jain & Pantel, 2010). Alpa Jain and Patrick Pantel. FactRank: Random Walks on a Web of Facts. In Proceedings of Conference on Computational Linguistics (COLING-10), 2010, pp. 501-509. Beijing, China

(Koo et al., 2010). Terry Koo, Alexander M. Rush, Michael Collins, Tommi Jaakkola, and David Sontag. Dual decomposition for parsing with non-projective head automata. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1288–1298, Cambridge, MA. Association for Computational Linguistics.

(Rush and Collins, 2011). Alexander M. Rush and Michael Collins. Exact Decoding of Syntactic Translation Models through Lagrangian Relaxation. In Proc. of the Annual Meeting of the ACL, 2011.

(Schank and Colby 1973). Roger C. Schank and Kenneth Mark Colby, editors. Computer models of thought and language. San Francisco, W. H. Freeman, 1973.

(Shavlik et. al., 2009). Jude Shavlik and Sriraam Natarajan. Speeding up Inference in Markov Logic Networks By Preprocessing to Reduce the Size of the Resulting Grounded Network. In Proc. of the International Joint Conference in Artificial Intelligence (IJCAI), 2009. http://ftp.cs.wisc.edu/machine-learning/shavlik-group/shavlik.ijcai09.pdf

(Weld et. al., 2008). Daniel S. Weld, Fei Wu, Eytan Adar, Saleema Amershi, James Fogarty, Raphael Hoffmann, Kayur Patel and Michael Skinner. Intelligence in Wikipedia. In Prov. Twenty-Third Conference on Artificial Intelligence (AAAI-08), Chicago, IL, July, 2008.

7. ACRONYMS

ACE	Attempto Controlled English
ALBP	Anytime Lifted Belief Propagation
BP	Belief Propagation
CAF	Common Annotation Format
CCG	Combinatory Categorial Grammar
CCM	Constrained Conditional Model
CSLI	Stanford's Center for the Study of Language and Information
DBLP	A computer science bibliography website hosted at Universität Trier, originally a database and logic programming bibliography site. http://en.wikipedia.org/wiki/DBLP
DD	Dual Decomposition
DSRS	Domain Specific Reasoning System (provided by the ET)
EEE	Event Extraction Experiment
EM	Expectation Maximization
EAT	Event Annotation Tool
ET	Evaluation Team, contractor selected by the Government
FAUST	Flexible Acquisition and Understanding System for Text, SRI's Machine Reading system
IC	Intelligence Community, name of a use case defined by the Government ET
IC++	An enhanced version of the IC use case for MRP
IE	Information Extraction
IHMC	Institute for Human and Machine Cognition
IID	Independent and Identically Distributed
ILP	Inductive Logic Programming
JI	Joint Inference
KBP	Knowledgebase Population
LBP	Lifted Belief Propagation
LDA	Latent Dirichlet Allocation
LDC	Linguistic Data Consortium, University of Pennsylvania, part of the ET
LIDE	Lifted Inference with Distinct Evidence

LP	Linear Programming
LVC	Light-Verb Constructions
MAP	Maximum *a Posteriori*
MCMC	Markov Chain Monte Carlo
MIML-RE	Multi-Instance, Multi-Label Relation Extraction
MLN	Markov Logic Networks
MR	Machine Reading
MRAPI	Machine Reading Application Programming Interface
MRP	Machine Reading Program
MR-KBP	Machine Reading Knowledgebase Population
MV-RNN	Matrix Vector Recursive Neural Net
NER	Named Entity Recognition
NFL	National Football League, name of a use case defined by the Government ET
NL	Natural Language
NLP	Natural Language Processing
NLU	Natural Language Understanding
OLPI	Ontological Lifted Probabilistic Inference
PCE	Probabilistic Consistency Engine
PCFG	Probabilistic Context-Free Grammar
QALD	Question Answering over Linked Data
RDN	Relational Dependency Networks
RE	Relation Extraction
RNN	Recursive Neural Network
SE	Software Engineering
SPN	Sum Product Network
SRL	Semantic Role-Labeling
SVM	Support Vector Machines
TAC-KBP	Text Analysis Conference Knowledgebase Population
TE	Textual Entailment
TREC	Text REtrieval Conference

UI	User Interface
UIUC	University of Illinois Urbana-Champaign
UMass	University of Massachusetts at Amherst
USP	Unsupervised Semantic Parsing, University of Washington
UW	University of Washinton
WILL	Wisconsin Inductive Logic Learner
WSD	Web Services for Devices
XFST	Xerox Finite State Tool

APPENDIX A. FAUST MACHINE READING PUBLICATIONS

The FAUST team made extensive contributions to scientific knowledge. Our world-leading researchers have already won five best paper awards and have published 155 papers, most in top conferences, for their papers on MR-sponsored and MR-related work. Numbered cites in this report refer to the table of publications below.

Table 1: FAUST Machine reading publications

1	Hoifung Poon, Pedro Domingos	*Unsupervised Semantic Parsing* (EMNLP 2009 Best paper award)	Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)	August 6–7, 2009	Singapore
2	Mark Sammons, V.G. Vinod Vydiswaran, Tim Vieira, Nikhil Johri, Ming-Wei Chang, Dan Goldwasser, Vivek Srikumar, Gourab Kundu, Yuancheng Tu, Kevin Small, Joshua Rule, Quang Do, Dan Roth	*Relation Alignment for Textual Entailment Recognition*	NIST Text Analysis Conference (TAC 2009)	November 16–17, 2009	Gaithersburg, Maryland
3	Michael Wick, Khashayar Rohanimanesh, Sameer Singh, Andrew McCallum	*Training Factor Graphs with Reinforcement Learning for Efficient MAP Inference* (Spotlight award)	24th Annual Conference on Neural Information Processing Systems (NIPS 2009)	December 7–10, 2009	Vancouver, British Columbia, Canada
4	Andrew McCallum, Karl Schultz, Sameer Singh	*FACTORIE: Probabilistic Programming via Imperatively Defined Factor Graphs* (Spotlight award)	24th Annual Conference on Neural Information Processing Systems (NIPS 2009)	December 7–10, 2009	Vancouver, British Columbia, Canada
5	Michael Wick, Khashayar Rohanimanesh, Aron Culotta, Andrew McCallum	*SampleRank: Learning Preferences from Atomic Gradients*	24th Annual Conference on Neural Information Processing Systems (NIPS 2009) Workshop on Advances in Ranking	December 11, 2009	Vancouver, British Columbia, Canada

6	Sriraam Natarajan, Prasad Tadepalli, Gautam Kunapuli, Jude Shavlik	*Learning Parameters for Relational Probabilistic Models with Noisy-Or Combining Rule*	8th International Conference on Machine Learning and Applications (ICMLA 2009)	December 13–15, 2009	Miami, Florida
7	Ming-Wei Chang, Dan Goldwasser, Dan Roth, Vivek Srikumar	*Discriminative Learning over Constrained Latent Representations*	11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2010)	June 1–6, 2010	Los Angeles, California
8	Valentin I. Spitkovsky, Daniel Jurafsky, Hiyan Alshawi	*Profiting from Mark-Up: Hyper-Text Annotations for Guided Parsing*	11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2010)	June 1–6, 2010	Los Angeles, California
9	Valentin I. Spitkovsky, Hiyan Alshawi, Daniel Jurafsky	*From Baby Steps to Leapfrog: How “Less is More” in Unsupervised Dependency Parsing*	11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2010)	June 1–6, 2010	Los Angeles, California
10	Mihai Surdeanu, Christopher D. Manning	*Ensemble Models for Dependency Parsing: Cheap and Good?*	11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2010)	June 1–6, 2010	Los Angeles, California

11	Annie Zaenen, Cleo Condoravdi, Daniel G. Bobrow, Raphael Hoffmann	Supporting rule-based representations with corpus-derived lexical information	11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2010)	June 1–6, 2010	Los Angeles, California
12	Sebastian Riedel, David A. Smith	Relaxed Marginal Inference and its Application to Dependency Parsing	11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2010)	June 1–6, 2010	Los Angeles, California
13	Sameer Singh, Limin Yao, Sebastian Riedel, Andrew McCallum	Constraint-Driven Rank-Based Learning for Information Extraction	11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2010)	June 1–6, 2010	Los Angeles, California
14	Ming-Wei Chang, Vivek Srikumar, Dan Goldwasser, Dan Roth	Structured Output Learning with Indirect Supervision	27th International Conference on Machine Learning (ICML 2010)	June 21–24, 2010	Haifa, Israel
15	Gregory Druck, Andrew McCallum	High-Performance Semi-Supervised Learning using Discriminatively Constrained Generative Models	27th International Conference on Machine Learning (ICML 2010)	June 21–24, 2010	Haifa, Israel
16	Sriraam Natarajan, Tushar Khot, Kristian Kersting, Bernd Guttmann, Jude Shavlik	Boosting Relational Dependency Networks	20th International Conference on Inductive Logic Programming (ILP 2010)	June 27–30, 2010	Firenze, Italy
17	Sebastian Riedel, David A. Smith, Andrew McCallum	Inference by Minimizing Size, Divergence, or their Sum	26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)	July 8–11, 2010	Catalina Island, California

18	Xiao Ling and Daniel S. Weld	Temporal Information Extraction	24th AAAI Conference on Artificial Intelligence (AAAI-10)	July 11–15, 2010	Atlanta, Georgia
19	Mark Sammons, V. G. Vinod Vydiswaran, Dan Roth	Ask not what Textual Entailment can do for you...	48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)	July 11–16, 2010	Uppsala, Sweden
20	Fei Wu, Daniel S. Weld	Open Information Extraction using Wikipedia	48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)	July 11–16, 2010	Uppsala, Sweden
21	Raphael Hoffmann, Congle Zhang, Dan Weld	Learning 5000 Relational Extractors	48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)	July 11–16, 2010	Uppsala, Sweden
22	Jenny Rose Finkel, Christopher D. Manning	Hierarchical Joint Learning: Improving Joint Parsing and Named Entity Recognition with Non-Jointly Labeled Data	48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)	July 11–16, 2010	Uppsala, Sweden
23	Sriraam Natarajan, Tushar Khot, Daniel Lowd, Prasad Tadepalli, Kristian Kersting, Jude Shavlik	Exploiting Causal Independence in Markov Logic Networks: Combining Undirected and Directed Models	24th AAAI Conference on Artificial Intelligence (AAAI-10) Workshop on Statistical Relational AI	July 12, 2010	Atlanta, Georgia
24	Sebastian Riedel	Declarative Probabilistic Programming for Undirected Models: Open Up to Scale Up	24th AAAI Conference on Artificial Intelligence (AAAI-10) Workshop on Statistical Relational AI	July 12, 2010	Atlanta, Georgia
25	Valentin I. Spitkovsky, Hiyan Alshawi, Daniel Jurafsky, Christopher D. Manning	Viterbi Training Improves Unsupervised Dependency Parsing	14th Conference on Computational Natural Language Learning (CoNLL-2010)	July 15–16, 2010	Uppsala, Sweden

26	James Clarke, Dan Goldwasser, Ming-Wei Chang, Dan Roth	Driving Semantic Parsing from the World's Response	14th Conference on Computational Natural Language Learning (CoNLL-2010)	July 15–16, 2010	Uppsala, Sweden
27	Yee Seng Chan, Dan Roth	Exploiting Background Knowledge for Relation Extraction	23rd International Conference on Computational Linguistics (COLING 2010)	August 23–27, 2010	Beijing, China
28	Cleo Condoravdi, Sven Lauer	Performative Verbs and Performative Acts	15th Sinn und Bedeutung Conference	September 9–11, 2010	Saarbrücken, Germany
29	Michael Wick, Andrew McCallum, Gerome Miklau	Scalable Probabilistic Databases with Factor Graphs and MCMC	36th International Conference on Very Large Data Bases (VLDB 2010)	September 13–17, 2010	Singapore
30	Sebastian Riedel, Limin Yao, Andrew McCallum	Modeling Relations and Their Mentions without Labeled Text	European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2010)	September 20–24, 2010	Barcelona, Spain
31	Sriraam Natarajan, Tushar Khot, Daniel Lowd, Prasad Tadepalli, Kristian Kersting, Jude Shavlik	Exploiting Causal Independence in Markov Logic Networks: Combining Undirected and Directed Models	European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2010)	September 20–24, 2010	Barcelona, Spain
32	Michael Connor, Ming-Wei Chang, Dan Roth	The Necessity of Combining Adaptation Methods	2010 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2010)	October 9–11, 2010	MIT, Massachusetts
33	Quang Xuan Do, Dan Roth	Relational Constraint-based Taxonomic Relation Classification	2010 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2010)	October 9–11, 2010	MIT, Massachusetts

34	Stefan Schoenmackers, Jesse Davis, Oren Etzioni, Daniel S. Weld	*Learning First-Order Horn Clauses from Web Text*	2010 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2010)	October 9–11, 2010	MIT, Massachusetts
35	Limin Yao, Sebastian Riedel, Andrew McCallum	*Collective Cross-Document Relation Extraction without Labeled Data*	2010 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2010)	October 9–11, 2010	MIT, Massachusetts
36	Terry Koo, Alexander M. Rush, Michael Collins, Tommi Jaakkola, David Sontag	*Dual Decomposition for Parsing with Non-Projective Head Automata (Best Paper Award)*	2010 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2010)	October 9–11, 2010	MIT, Massachusetts
37	Alexander M. Rush, David Sontag, Michael Collins, Tommi Jaakkola	*On Dual Decomposition and Linear Programming Relaxations for Natural Language Processing*	2010 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2010)	October 9–11, 2010	MIT, Massachusetts
38	Mihai Surdeanu, David McClosky, Julie Tibshirani, John Bauer, Angel Chang, Valentin I. Spitkovsky, Christopher D. Manning	*A Simple Distant Supervision Approach for the KBP Slot Filling Task*	NIST Text Analysis Conference (TAC 2010)	November 15–16, 2010	Gaithersburg, Maryland
39	Angel X. Chang, Valentin I. Spitkovsky, Eric Yeh, Eneko Agirre, Christopher D. Manning	*Stanford-UBC Entity Linking at TAC-KBP*	NIST Text Analysis Conference (TAC 2010)	November 15–16, 2010	Gaithersburg, Maryland
40	Richard Socher, Christopher D. Manning, Andrew Y. Ng	*Learning Continuous Phrase Representations and Syntactic Parsing with Recursive Neural Networks*	24th Annual Conference on Neural Information Processing Systems (NIPS 2010) Deep Learning and Unsupervised Feature Learning Workshop	December 10, 2010	Whistler, British Columbia, Canada

41	Ramesh Nallapati, Mihai Surdeanu, Christopher Manning	*Blind domain transfer for Named Entity Recognition using Generative Latent Topic Models*	24th Annual Conference on Neural Information Processing Systems (NIPS 2010) Workshop on Transfer Learning using Rich Generative Models	December 10, 2010	Whistler, British Columbia, Canada
42	Abhay Jha, Vibhav Gogate, Alexandra Meliou, Dan Suciu	*Lifted Inference from the Other Side: The tractable Features*	24th Annual Conference on Neural Information Processing Systems (NIPS 2010)	December 10–11, 2010	Whistler, British Columbia, Canada
43	Vibhav Gogate, William Austin Webb, Pedro Domingos	*Learning Efficient Markov networks*	24th Annual Conference on Neural Information Processing Systems (NIPS 2010)	December 10–11, 2010	Whistler, British Columbia, Canada
44	Sameer Singh, Amarnag Subramanya, Fernando Pereira, Andrew McCallum	*Distributed MAP Inference for Undirected Graphical Models*	24th Annual Conference on Neural Information Processing Systems (NIPS 2010) Workshop on Learning on Cores, Clusters and Clouds	December 10–11, 2010	Whistler, British Columbia, Canada
45	Cleo Condoravdi	*NPI licensing in temporal clauses*	Natural Language and Linguistic Theory, Vol. 28.4	December 18, 2010	(journal article)
46	Mark Sammons, V.G.Vinod Vydiswaran, Dan Roth	*Recognizing Textual Entailment*	A Chapter in "Multilingual Natural Language Applications: From Theory to Practice (2011)"	2011	(book chapter)
47	M. Levent Koc, Christopher Re	*Incrementally Maintaining Classification using an RDBMS*	Proceedings of the VLDB Endowment, Vol. 4, No. 5	February 2011	N/A
48	Feng Niu, Christopher Re, AnHai Doan, Jude Shavlik	*Tuffy: Scaling up Statistical Inference in Markov Logic Networks using an RDBMS*	Proceedings of the VLDB Endowment, Vol. 4, No. 6	March 2011	N/A

49	Fei Chen, Xixuan Feng, Christopher Ré, Min Wang	Optimizing Statistical Information Extraction Programs Over Evolving Text	28th IEEE International Conference on Data Engineering (ICDE 2012)	April 1–5, 2012	Arlington, Virginia
50	Dan Suciu, Dan Olteanu, Christopher Ré, Christoph Koch	Probabilistic Databases	Book published by Morgan and Claypool	May 2011	(book)
51	Mihai Surdeanu, Massimiliano Ciaramita, Hugo Zaragoza	Learning to Rank Answers to Non-Factoid Questions from Web Collections	Computational Linguistics 37(2)	June 2011	(journal article)
52	David McClosky, Mihai Surdeanu, Christopher D. Manning	Event Extraction as Dependency Parsing	49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)	June 19–24, 2011	Portland, Oregon
53	Nathanael Chambers, Dan Jurafsky	Template-Based Information Extraction without the Templates	49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)	June 19–24, 2011	Portland, Oregon
54	John Lee, Jason Naradowsky, David A. Smith	A Discriminative Model for Joint Morphological Disambiguation and Dependency Parsing	49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)	June 19–24, 2011	Portland, Oregon
55	Sameer Singh, Amarnag Subramanya, Fernando Pereira, Andrew McCallum	Large-Scale Cross-Document Coreference Using Distributed Inference and Hierarchical Models	49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)	June 19–24, 2011	Portland, Oregon

56	Rafael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, Dan Weld	*Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations*	49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)	June 19–24, 2011	Portland, Oregon
57	Lev Ratinov, Dan Roth, Doug Downey, Mike Anderson	*Local and Global Algorithms for Disambiguation to Wikipedia*	49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)	June 19–24, 2011	Portland, Oregon
58	Yee Seng Chan, Dan Roth	*Exploiting Syntactico-Semantic Structures for Relation Extraction*	49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)	June 19–24, 2011	Portland, Oregon
59	Dan Goldwasser, Roi Reichart, James Clarke, Dan Roth	*Confidence Driven Unsupervised Semantic Parsing*	49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)	June 19–24, 2011	Portland, Oregon
60	Yee Seng Chan, Dan Roth (duplicates 58)	*Exploiting Syntactico-Semantic Structures for Relation Extraction*	49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)	June 19–24, 2011	Portland, Oregon
61	Yuancheng Tu, Dan Roth	*Learning English Light Verb Constructions: Contextual or Statistical*	49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011) Workshop on Multiword Expressions	June 23, 2011	Portland, Oregon
62	Gourab Kundu, Dan Roth	*Adapting Text instead of the Model: An Open Domain Approach*	15th Conference on Computational Natural Language Learning (CoNLL-2011)	June 23–24, 2011	Portland, Oregon

63	Kai-Wei Chang, Rajhans Samdani, Alla Rozovskaya, Nick Rizzolo, Mark Sammons, Dan Roth	*Inference Protocols for Co-reference Resolution*	15th Conference on Computational Natural Language Learning (CoNLL-2011)	June 23–24, 2011	Portland, Oregon
64	Gourab Kundu, Dan Roth	*Adapting Text Instead of the Model: An Open Domain Approach* (Best student paper award)	15th Conference on Computational Natural Language Learning (CoNLL-2011)	June 23–24, 2011	Portland, Oregon
65	Sebastian Riedel, Andrew McCallum	*Robust Biomedical Event Extraction with Dual Decomposition and Minimal Domain Adaptation*	49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011) Workshop on Biomedical Natural Language Processing (BioNLP 2011)	June 24, 2011	Portland, Oregon
66	David McClosky, Mihai Surdeanu, Christopher D. Manning	*Event Extraction as Dependency Parsing in BioNLP 2011*	49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011) Workshop on Biomedical Natural Language Processing (BioNLP 2011)	June 24, 2011	Portland, Oregon
67	Sebastian Riedel, David McClosky, Mihai Surdeanu, Andrew McCallum, Christopher D. Manning	*Model Combination for Event Extraction in BioNLP 2011*	49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011) Workshop on Biomedical Natural Language Processing (BioNLP 2011)	June 24, 2011	Portland, Oregon

68	Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky	Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task.	15th Conference on Computational Natural Language Learning (CoNLL-2011) Shared Task	June 24, 2011	Portland, Oregon
69	Mihai Surdeanu, David McClosky, Mason R. Smith, Andrey Gusev, Christopher D. Manning	Customizing an Information Extraction System to a New Domain	49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011) Workshop on Relational Models of Semantics	June 24, 2011	Portland, Oregon
70	Richard Socher, Cliff Lin, Andrew Y. Ng, Christopher D. Manning	Parsing Natural Scenes and Natural Language with Recursive Neural Networks (Distinguished Paper Award)	28th International Conference on Machine Learning (ICML 2011)	June 28, 2011	Seattle, Washington
71	Gourab Kundu, Ming-Wei Chang, Dan Roth	Prior Knowledge Driven Domain Adaptation	28th International Conference on Machine Learning (ICML 2011) Workshop on Combining Learning Strategies to Reduce Label Cost	June 28 – July 2, 2011	Bellevue, Washington
72	Hoifung Poon, Pedro Domingos	A New Deep Architecture	27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)	July 14–17, 2011	Barcelona, Spain
73	Vibhav Gogate, Pedro Domingos	Probabilistic Theorem Proving	27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)	July 14–17, 2011	Barcelona, Spain
74	Vibhav Gogate, Pedro Domingos	Approximation by Quantization	27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)	July 14–17, 2011	Barcelona, Spain

75	Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, Christopher D. Manning	*Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions*	Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)	July 27–31, 2011	Edinburgh, Scotland, UK
76	Sebastian Riedel, Andrew McCallum	*Fast and Robust Joint Models for Biomedical Event Extraction*	Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)	July 27–31, 2011	Edinburgh, Scotland, UK
77	Limin Yao, Aria Haghighi, Sebastian Riedel, Andrew McCallum	*Structured Relation Discovery using Generative Models*	Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)	July 27–31, 2011	Edinburgh, Scotland, UK
78	Chloé Kiddon, Pedro Domingos	*Coarse-to-Fine Inference and Learning for First-Order Probabilistic Models*	25th Conference on Artificial Intelligence (AAAI-11)	August 7–11, 2011	San Francisco, California
79	Gourab Kundu, Ming-Wei Chang, Dan Roth	*Prior Knowledge Driven Domain Adaptation*	28th International Conference on Machine Learning (ICML 2011), Workshop on Combining Learning Strategies to Reduce Label Cost	July 2, 2011	Bellevue, Washington
80	Quang Do, Yee Seng Chan, Dan Roth	*Minimally Supervised Event Causality Extraction*	Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)	July 27–29, 2011	Edinburgh, Scotland, UK
81	Vivek Srikumar, Dan Roth	*A Joint Model for Extended Semantic Role Labeling*	Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)	July 27–29, 2011	Edinburgh, Scotland, UK

82	Kai-Wei Chang, Dan Roth	Selective Block Minimization for Faster Convergence of Limited Memory Large-scale Linear Models	17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2011)	August 21–24, 2011	San Diego, California
83	Dan Goldwasser, Dan Roth	Learning from Natural Instructions	22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)	July 16–22, 2011	Barcelona, Spain
84	Hoifung Poon, Pedro Domingos	Sum-Product Networks: A New Deep Architecture	27th Conference on Uncertainty in Artificial Intelligence (UAI)	July 14–17, 2011	Barcelona, Spain
85	Feng Niu, Ce Zhang, Christopher Ré, Jude Shavlik	Felix: Scaling Inference for Markov Logic with an Operator-based Approach	arXiv e-prints	August 1, 2011	(arXiv)
86	Mehmet Levent Koc, Christopher Ré	Incrementally Maintaining Classification using an RDBMS	37th International Conference on Very Large Data Bases (VLDB 2011)	August 29 – September 3, 2011	Seattle, Washington
87	Feng Niu, Christopher Ré, Anhai Doan, Jude Shavlik	Tuffy: Scaling up Statistical Inference in Markov Logic Networks using an RDBMS	37th International Conference on Very Large Data Bases (VLDB 2011)	August 29 – September 3, 2011	Seattle, Washington
88	Shalini Ghosh, Natarajan Shankar, Sam Owre	Machine Reading Using Markov Logic Networks for Collective Probabilistic Inference	European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML PKDD 2011) Workshop on Collective Learning and Inference from Structured data (CoLISD 2011)	September 9, 2011	Athens, Greece

89	Mihai Surdeanu, Sonal Gupta, John Bauer, David McClosky, Angel X. Chang, Valentin I. Spitkovsky, Christopher D. Manning	*Stanford's Distantly-Supervised Slot-Filling System*	Text Analysis Conference Knowledgebase Population Workshop (TAC KBP 2011)	November 14–15, 2011	Gaithersburg, Maryland
90	Angel X. Chang, Valentin I. Spitkovsky, Eneko Agirre, Christopher D. Manning	*Stanford-UBC Entity Linking at TAC-KBP, Again*	Text Analysis Conference Knowledgebase Population Workshop (TAC KBP 2011)	November 14–15, 2011	Gaithersburg, Maryland
91	Lev Ratinov, Dan Roth	*GLOW TAC-KBP 2011 Entity Linking System*	Text Analysis Conference Knowledgebase Population Workshop (TAC KBP 2011)	November 14–15, 2011	Gaithersburg, Maryland
92	Christopher Ré, Dan Suciu	*Understanding Cardinality Estimation using Entropy Maximization*	ACM Transactions on Database Systems (TODS), 2012:37(1)	December 2011	(journal article)
93	Tushar Khot, Sriraam Natarajan, Kristian Kersting, Jude Shavlik	*Learning Markov Logic Networks via Functional Gradient Boosting*	2011 IEEE International Conference on Data Mining (ICDM 2011)	December 11–14, 2011	Vancouver, British Columbia, Canada
94	Feng Niu, Benjamin Recht, Christopher Ré, Stephen J. Wright	*Hogwild!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent*	25th Annual Conference on Neural Information Processing Systems (NIPS 2011)	December 13–15, 2011	Granada, Spain
95	Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, Christopher D. Manning	*Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection*	25th Annual Conference on Neural Information Processing Systems (NIPS 2011)	December 13–15, 2011	Granada, Spain

96	Sameer Singh, Andrew McCallum	Toward Asynchronous Distributed MCMC Inference for Large Graphical Models (Spotlight award)	Neural Information Processing Systems (NIPS 2011), Big Learning Workshop on Algorithms, Systems, and Tools for Learning at Scale	December 16–17, 2011	Sierra Nevada, Spain
97	Sameer Singh, Brian Martin, Andrew McCallum	Inducing Value Sparsity for Parallel Inference in Tree-shaped Models	Neural Information Processing Systems (NIPS 2011), Workshop on Computational trade-offs in Statistical Learning (COST)	December 16–17, 2011	Sierra Nevada, Spain
98	Sriraam Natarajan, Tushar Khot, Kristian Kersting, Bernd Gutmann, Jude Shavlik	Gradient-based Boosting for Statistical Relational Learning: The Relational Dependency Network Case	Machine Learning Journal (volume 86, number 1, pp. 25–56)	January 2012	(journal article)
99	Quang Do, Dan Roth	Exploring the Wikipedia Structure in Local and Global Classification of Taxonomic Relations	Journal of Natural Language Engineering (JNLE)	April 2012	(journal article)
100	Fei Chen, Xixuan Feng, Christopher Re, Min Wang	Optimizing Statistical Information Extraction Programs Over Evolving Text	28th IEEE International Conference on Data Engineering (ICDE 2012)	April 1–5, 2012	Arlington, Virginia
101	Angel X. Chang, Christopher D. Manning	SUTime: A Library for Recognizing and Normalizing Time Expressions	8th International Conference on Language Resources and Evaluation (LREC 2012)	May 23–25, 2012	Istanbul, Turkey

102	Valentin I. Spitkovsky, Angel X. Chang	*A Cross-Lingual Dictionary for English Wikipedia Concepts*	8th International Conference on Language Resources and Evaluation (LREC 2012)	May 23–25, 2012	Istanbul, Turkey
103	James Clarke, Vivek Srikumar, Mark Sammons, Dan Roth	*An NLP Curator (or: How I Learned to Stop Worrying and Love NLP Pipelines)*	8th International Conference on Language Resources and Evaluation (LREC 2012)	May 23–25, 2012	Istanbul, Turkey
104	David McClosky, Sebastian Riedel, Mihai Surdeanu, Andrew McCallum, Christopher D. Manning	*Combining Joint Models for Biomedical Event Extraction*	BMC Bioinformatics	June 2012	(journal article)
105	Ming-Wei Chang, Lev Ratinov, Dan Roth	*Structured Learning with Constrained Conditional Models*	Journal of Machine Learning	June 2012	(journal article)
106	Gabor Angeli, Christopher D. Manning, Daniel Jurafsky	*Parsing Time: Learning to Interpret Time Expressions*	The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2012)	June 3–8, 2012	Montréal, Canada
107	Valentin I. Spitkovsky, Hiyan Alshawi, Daniel Jurafsky	*Capitalization Cues Improve Dependency Grammar Induction*	The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2012)	June 3–8, 2012	Montréal, Canada
108	R Zhao Quang Do, Dan Roth	*A Robust Shallow Temporal Reasoning System*	The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2012)	June 3–8, 2012	Montréal, Canada

109	Rajhans Samdani, Ming-Wei Chang, Dan Roth	*Unified Expectation Maximization*	The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2012)	June 3–8, 2012	Montreal, Canada
110	Yuancheng Tu, Dan Roth	*Sorting out the Most Confusing English Phrasal Verbs*	Association for Computational Linguistics First Joint Conference on Lexical and Computational Semantics	June 7–8, 2012	Montréal, Canada
111	Lauri Karttunen	*Simple and Phrasal Implicatives*	Association for Computational Linguistics First Joint Conference on Lexical and Computational Semantics	June 7–8, 2012	Montréal, Canada
112	Limin Yao, Sebastian Riedel, Andrew McCallum	*Probabilistic Databases of Universal Schema*	2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2012) Joint Workshop on Automatic Knowledgebase Construction and Web-scale Knowledge Extraction (AKBC-WEKEX 2012)	June 7–8, 2012	Montréal, Canada
113	Rajhans Samdani, Dan Roth	*Efficient Decomposed Learning for Structured Prediction*	29th International Conference on Machine Learning (ICML 2012)	June 26 – July 1, 2012	Edinburgh, Scotland, UK

114	David Belanger, Alexandre Passos, Sebastian Riedel, Andrew McCallum	Speeding up MAP with Column Generation and Block Regularization	29th International Conference on Machine Learning (ICML 2012) Workshop on Inference: Interactions between Inference and Learning	June 30, 2012	Edinburgh, Scotland, UK
115	Rajhans Samdani, Ming-Wei Chang, Dan Roth	A Framework for Tuning Posterior Entropy in Unsupervised Learning	29th International Conference on Machine Learning (ICML 2012) Workshop on Inference: Interactions between Inference and Learning	June 30, 2012	Edinburgh, Scotland, UK
116	Tushar Khot, Sriraam Natarajan, Kristian Kersting, Jude Shavlik	Structure Learning with Hidden Data in Relational Domains	29th International Conference on Machine Learning (ICML 2012) Statistical Relational Learning Workshop	June 30, 2012	Edinburgh, Scotland, UK
117	Wanxiang Che, Valentin I. Spitkovsky, Ting Liu	A Comparison of Chinese Parsers for Stanford Dependencies	50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)	July 8–14, 2012	Jeju Island, South Korea
118	Eric H. Huang, Richard Socher, Christopher D. Manning, Andrew Y. Ng	Improving Word Representations via Global Context and Multiple Word Prototypes	50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)	July 8–14, 2012	Jeju Island, South Korea
119	Shay Cohen, Karl Stratos, Michael Collins, Dean Foster, Lyle Ungar	Spectral learning of Latent-Variable PCFGs	50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)	July 8–14, 2012	Jeju Island, South Korea
120	Wei Lu, Dan Roth	Automatic Event Extraction with Structured Preference Modeling	50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)	July 9–11, 2012	Jeju Island, South Korea

121	Ce Zhang, Feng Niu, Christopher Re, Jude Shavlik	Big Data versus the Crowd: Looking for Relationships in All the Right Places	50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)	July 9–11, 2012	Jeju Island, South Korea
122	Jason Naradowsky, Sebastian Riedel, David A. Smith	Improving NLP through Marginalization of Hidden Syntactic Structure	Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)	July 12–14, 2012	Jeju Island, South Korea
123	Sebastian Riedel, David A. Smith, Andrew McCallum	Parse, Price and Cut - Delayed Column and Row Generation for Graph Based Parsers	Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)	July 12–14, 2012	Jeju Island, South Korea
124	Richard Socher, Brody Huval, Christopher D. Manning, Andrew Y. Ng	Semantic Compositionality through Recursive Matrix-Vector Spaces	Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)	July 12–14, 2012	Jeju Island, South Korea
125	David McClosky, Christopher D. Manning	Learning Constraints for Consistent Timeline Extraction	Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)	July 12–14, 2012	Jeju Island, South Korea
126	Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, Dan Jurafsky	Joint Entity and Event Coreference Resolution across Documents	Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)	July 12–14, 2012	Jeju Island, South Korea

127	Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, Christopher D. Manning	Multi-instance Multi-label Learning for Relation Extraction	Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)	July 12–14, 2012	Jeju Island, South Korea
128	Mengqiu Wang, Christopher D. Manning	Probabilistic Finite State Machines for Regression-based MT Evaluation	Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)	July 12–14, 2012	Jeju Island, South Korea
129	Valentin I. Spitkovsky, Hiyan Alshawi, Daniel Jurafsky	Three Dependency-and-Boundary Models for Grammar Induction	Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)	July 12–14, 2012	Jeju Island, South Korea
130	Lev Ratinov, Dan Roth	Learning-based Multi-Sieve Co-Reference Resolution with Knowledge	Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)	July 12–14, 2012	Jeju Island, South Korea
131	Vivek Srikumar, Gourab Kundu, Dan Roth	On Amortizing Inference Cost for Structured Prediction	Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)	July 12–14, 2012	Jeju Island, South Korea

132	Quang Do, Wei Lu, Dan Roth	*Joint Inference for Event Timeline Construction*	Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)	July 12–14, 2012	Jeju Island, South Korea
133	Kai-Wei Chang, Rajhans Samdani, Alla Rozovskaya, Mark Sammons, Dan Roth	*Illinois-Coref: The UI System in the CoNLL-2012 Shared Task*	Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012) CoNLL Shared Task	July 13, 2012	Jeju Island, South Korea
134	Juan F. Mancilla-Caceres, Wen Pu, Dorothy Espelage, Eyal Amir	*Identifying Bullies with a Computer Game*	26th Conference on Artificial Intelligence (AAAI-12)	July 22–26, 2012	Toronto, Canada
135	Mark Richards, Eyal Amir	*Information-Set Generation in Partially Observable Games*	26th Conference on Artificial Intelligence (AAAI-12)	July 22–26, 2012	Toronto, Canada
136	Xiao Ling and Daniel S. Weld	*Fine-Grained Entity Recognition*	26th Conference on Artificial Intelligence (AAAI-12)	July 22–26, 2012	Toronto, Canada
137	Congle Zhange, Raphael Hoffmann, Daniel S. Weld	*Ontological Smoothing for Relation Extraction with Minimal Supervision*	26th Conference on Artificial Intelligence (AAAI-12)	July 22–26, 2012	Toronto, Canada
138	Hung Bui, Tuyen Huynh, Rodrigo de Salvo Braz	*Exact Lifted Inference with Distinct Soft Evidence on Every Object*	26th Conference on Artificial Intelligence (AAAI-12)	July 22–26, 2012	Toronto, Canada
139	Rodrigo de Salvo Braz, Shahin Saadati, Hung Bui, Ciaran O'Reilly	*Lifted Arbitrary Constraint Solving for Lifted Probabilistic Inference*	Uncertainty in Artificial Intelligence Conference (UAI 2012) 2nd International Workshop on Statistical Relational AI	August 15–18, 2012	Avalon, California

140	Hung Bui, Tuyen Huynh, Sebastian Riedel	*Automorphism Groups of Graphical Models and Lifted Variational Inference*	Uncertainty in Artificial Intelligence Conference (UAI 2012) 2nd International Workshop on Statistical Relational AI	August 15–18, 2012	Avalon, California
141	Richard G. Freedman, Rodrigo de Salvo Braz, Hung Bui, Sriraam Natarajan	*Initial Empirical Evaluation of Anytime Lifted Belief Propagation*	Uncertainty in Artificial Intelligence Conference (UAI 2012) 2nd International Workshop on Statistical Relational AI	August 15–18, 2012	Avalon, California
142	Tushar Khot, Siddharth Srivastava, Sriraam Natarajan, Jude Shavlik	*Learning Relational Structure for Temporal Relation Extraction*	Uncertainty in Artificial Intelligence Conference (UAI 2012) 2nd International Workshop on Statistical Relational AI	August 15–18, 2012	Avalon, California
143	Feng Niu, Ce Zhang, Christopher Ré, Jude Shavlik	*DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference*	Very Large Database Search (VLDS 2012) The Second International Workshop on Searching and Integrating New Web Data Sources	August 31, 2012	Istanbul, Turkey
144	Codruta L. Girlea, Eyal Amir	Probabilistic Region Connection Calculus (Best paper award for the workshop)	European Conference on Artificial Intelligence (ECAI 2012) Workshop on Spatio-Temporal Dynamics (STeDy 2012)	August 27–28, 2012	Paris, France
145	Feng Niu, Ce Zhang, Christopher Ré, Jude Shavlik	*Elementary: Large-scale Knowledge-base Construction via Machine Learning and Statistical Inference*	International Journal on Semantic Web and Information Systems – Special Issue on Web-Scale Knowledge Extraction	September 2012	(journal article)

146	Valentin I. Spitkovsky, Hiyan Alshawi, Daniel Jurafsky	*Bootstrapping Dependency Grammar Inducers from Incomplete Sentence Fragments via Austere Models*	11th International Conference on Grammatical Inference (ICGI 2012)	September 12–15, 2012	College Park, Maryland
147	Gautam Kunapuli, Jude Shavlik	*Mirror Descent for Metric Learning: A Unified Approach*	European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2012)	September 24–28, 2012	Bristol, United Kingdom
148	Lauri Karttunen	*You Will Be Lucky to Break Even*	From Quirky Case to Representing Space: Papers in Honor of Annie Zaenen edited by Tracy Holloway King and Valeria de Paiva	December 2012	CSLI Publications, Stanford, California
149	Richard Socher, Brody Huval, Bharath Bhat, Christopher D. Manning, Andrew Y. Ng	*Convolutional-Recursive Deep Learning for 3D Object Classification*	26th Annual Conference on Neural Information Processing Systems (NIPS 2012)	December 3–6, 2012	Lake Tahoe, Nevada
150	David Belanger, Alexandre Passos, Sebastian Riedel, Andrew McCallum	*MAP Inference in Chains using Column Generation*	26th Annual Conference on Neural Information Processing Systems (NIPS 2012)	December 3–6, 2012	Lake Tahoe, Nevada
151	Marjorie McShane, Segei Nirenburg, Stephen Beale, Ben Johnson	*Resolving Elided Scopes of Modality in OntoAgent*	First Annual Conference on Advances in Cognitive Systems	December 6–8, 2012	Palo Alto, California
152	Jason Naradowsky, Tim Vieira, David A. Smith	*Grammarless Parsing for Joint Inference*	24th International Conference on Computational Linguistics (COLING 2012)	December 8–15, 2012	Mumbai, India
153	Feng Niu, Ce Zhang, Christopher Ré, Jude Shavlik	*Scaling Inference for Markov Logic via Dual Decomposition*	IEEE International Conference on Data Mining (ICDM 2012)	December 10–13, 2012	Brussels, Belgium

154	Cleo Condoravdi, Sven Lauer	*Meaning and Illocutionary Force*	Empirical Issues in Syntax and Semantics 9, edited by Christopher Piñón	December 31, 2012	Paris, France
155	Shay B. Cohen, Karl–Stratos, Michael Collins, Dean P. Foster, Lyle Ungar	*Experiments with spectral learning of latent-variable PCFGs*	2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NACCL HLT 2013)	June 10–12, 2013	Atlanta, Georgia
156	Pedro Domingos and W. Austin Webb	*A Tractable First-Order Probabilistic Logic*	26th Conference on Artificial Intelligence (AAAI-12)	July 22–26, 2012	Toronto, Canada

APPENDIX B. FAUST SOFTWARE MODULES

Table 2: FAUST Software Modules

Module Name; Source	Description	Download Instructions	Restrictions
RR & JI			
Alchemy 2.0; University of Washington	Software package for inference and learning in Markov Logic Networks (MLNS) that includes several lifted probabilistic inference algorithms.	http://code.google.com/p/alchemy-2/	MIT License
Unsupervised Semantic Parsing (USP); University of Washington	An algorithm for unsupervised semantic parsing that is now close to online and more scalable.	http://alchemy.cs.washington.edu/papers/poon09/	Modified BSD License
Sum Product Networks (SPNs); University of Washington	A deep architecture that is more general than arithmetic circuits and enables efficient exact inference. SPNs are directed acyclic graphs with variables as leaves, sums and products as internal nodes, and weighted edges.	http://alchemy.cs.washington.edu/spn/	Modified BSD License
Tuffy; University of Wisconsin	Tuffy is a state-of-the-art, highly-scalable, open-source Markov Logic Network inference engine that can perform efficient inference on very large data sets by utilizing the power of RDBMS.	http://hazy.cs.wisc.edu/hazy/tuffy/	GPL v3

Felix; University of Wisconsin	Felix is a relational optimizer that utilizes operator-based task decompositon to identify specialized subtasks in an MLN and decompose it effectively to ensure consistent semantics in terms of joint inference. Felix has Tuffy inside. Felix supports three specialized operators: classification, labeling and co-reference resolution, and one generic operator: MLN inference.	http://hazy.cs.wisc.edu/hazy/felix/	GPL v3
FACTORIE; University of Massachusetts	FACTORIE is a toolkit for deployable probabilistic modeling, implemented as a software library in Scala. It provides its users with a succinct language for creating relational factor graphs, estimating parameters and performing inference.	http://factorie.cs.umass.edu/	Apache 2 License
AIC Util; SRI	Utilities for AIC Expresso and AIC Praise projects that: extend the Google Guava libraries; provides utilities around String manipulation,Collections, Error,Math and Logging handling; implement a concurrency API to simplify branch & merging computation tasks; extending the Iterator semantics; and provide a configuration API.	https://code.google.com/p/aic-util/	BSD License
AIC Expresso; SRI	Symbolic manipulation/evaluation of expressions	https://code.google.com/p/aic-expresso/	BSD License
AIC Praise; SRI	Probabilistic Reasoning as Symbolic Evaluation framework	https://code.google.com/p/aic-praise/	BSD License
AIC Web Praise;SRI	AIC Praise Web Front-End	http://code.google.com/p/aic-web-praise/	BSD License

NLP			
Stanford CoreNLP; Stanford University	An integrated suite of natural language processing tools for English in Java, including tokenization, part-of-speech tagging, named entity recognition, parsing, and coreference.	http://nlp.stanford.edu/software/corenlp.shtml	GPL v2+
Stanford Biomedical Event Parser (SBEP); Stanford	Biomedical Event Extraction for the BioNLP 2009/2011 shared task.	http://nlp.stanford.edu/software/eventparser.shtml	GPL v2+
Stanford TokensRegex; Stanford - University	A tool for matching regular expressions over token sequences.	http://nlp.stanford.edu/software/tokensregex.shtml	GPL v2+
Stanford Temporal Tagger; Stanford University	Known as SUTime. A rule-based temporal tagger for English text.	http://nlp.stanford.edu/software/sutime.shtml	GPL v2+
Stanford Deterministic Coreference Resolution; Stanford University	This system implements the multi-pass sieve co-reference resolution (or anaphora resolution) system which won the CoNLL 2011 Shared Task and is described in Raghunathan et al. (EMNLP 2010).	http://nlp.stanford.edu/software/dcoref.shtml	GPL v2+

Illinois Curator; UIUC	The Illinois Curator is a configurable manager that provides a programmatic interface to distributed NLP components. The main Curator process is a service listening on a user-specified port; clients call the service with text to annotate and a component type, and the Curator handles the call to the relevant component plus any dependencies that must be satisfied (for example, Curator may automatically call a Part-of-Speech tagger first.) The NLP components are themselves run as services. This significantly simplifies development of complex NLP systems by allowing the same infrastructure to be used. UIUC is presently developing a distribution using Virtual Machines; when completed, this will replace the existing distribution.	http://cogcomp.cs.illinois.edu/page/software_view/Curator	Available to government and researchers under the Illinois free academic use license
Illinois Named Entity Recognizer; UIUC	The Illinois Named Entity Recognizer gives better, more robust performance across different domains, and supports a second, extended tag set (18 entity types), which will support event extraction even more effectively than the basic NER system.	http://cogcomp.cs.illinois.edu/page/software_view/NETagger	Available to government and researchers under the Illinois free academic use license

Illinois Wikifier; UIUC	The Illinois Wikifier is a concept-detection system that identifies sets of coherent concepts in open-domain text, identifying the WikiPedia page that most closely corresponds to them. It behaves somewhat like a cross-document co-reference system, though it only disambiguates proper nouns (not pronouns), and will link descriptive noun phrases to the relevant category page. Because it uses knowledge that is automatically extracted from Wikipedia's link structure and document text, it is remarkably robust across domains.	http://cogcomp.cs.illinois.edu/page/software_view/Wikifier	Available to government and researchers under the Illinois free academic use license
Illinois Semantic Role Labeler; UIUC	The Illinois Semantic Role Labeler identifies argument structure for every verb and deverbal noun in input text -- namely, who did what to whom. This tool provides a useful abstraction over underlying syntactic variations and is used in a number of UIUC’s more advanced NLP systems, such as the Illinois Event Extraction System and Illinois Coreference System.	http://cogcomp.cs.illinois.edu/page/software_view/SRL	Available to government and researchers under the Illinois free academic use license

Illinois Coreference System; UIUC	The Illinois Coreference System identifies chains of textual references to individual entities in free text, linking pronouns, descriptive phrases, and proper nouns. This capacity is a key requirement for systems performing textual inference, such as the Illinois Event Extraction System.	http://cogcomp.cs.illinois.edu/page/software_view/Coref	Available to government and researchers under the Illinois free academic use license
Illinois Temporal Reasoning System; UIUC	The Illinois Temporal Reasoning System identifies phrases relating to times, whether canonical dates or relative temporal expressions, and relates them to a reference time (which could be a document timestamp); all such expressions are normalized to a common data format. This capability is a key requisite for determining discourse structure, as it constrains relations between events.	http://cogcomp.cs.illinois.edu/page/software_view/IllinoisTemporalExtractor	Available to government and researchers under the Illinois free academic use license
Infobox Extractor; Wake Forest University	Creates Prolog facts from Wiki infoboxes. Contact: Sriraam Natarajan <snataraj@wakehealth.edu>	http://wfuhs.arane.us/MachineReading/InfoExtractor/	No license
Initial Long Range Example Creator; Wake Forest University	Java-based proof-of-concept demonstrating a technique for extracting information about entities--primarily people and organizations--within a corpus of documents.	http://corpus-query.wfuhs.arane.us/index.html	No license

Machine Learning (ML)

MultiR; University of Washington	A distantly supervised information extraction system, described in "Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations" [56].	http://www.cs.washington.edu/ai/raphaelh/mr	MultiR License Agreement (BSD like)
Fine-Grained Entity Recognition; University of Washington	Known as FIGER. Described in "Fine-Grained Entity Recognition" [136]. Download both the system and the training data.	http://www.cs.washington.edu/ai/figer/	MultiR License Agreement (BSD like)
RDN-Boost; University of Wisconsin	RDN-Boost implements Natarajan et al's scalable, gradient-based boosting algorithm structure learning for relational dependency networks (RDNs).	http://ftp.cs.wisc.edu/machine-learning/shavlik-group/WILL/rdnboost	GPL v3
Booster; University of Wisconsin	Booster implements Khot et al's scalable, gradient-based boosting algorithm for structure learning for Markov logic networks (MLNs).	http://ftp.cs.wisc.edu/machine-learning/shavlik-group/WILL/Boostr	GPL v3

Miscellaneous

SRI Gazetteer; SRI	Fast query and lookup for geopolitical entities according to the extensive datasource provided by http://www.geonames.org/	Contact Lynn Voss <loren.voss@sri.com>	Apache 2 License
Lexicon of Event Nominals; CSLI	Lexicon of Event Nominals	http://www.stanford.edu/group/csli_faust/Lexical_Resources/Event-Nominals/	Copyright 2012 CSLI
Lexicon of Verb Polarity; CSLI	Lexicon of very polarity including simple factive verbs, simple implicative verbs, and a lexicon of temporal dependencies in infinitivals (polarity plus time).	http://www.stanford.edu/group/csli_faust/Lexical_Resources/Polarity-Lexicon-of-Verbs/	Copyright 2012 CSLI

Lexicon of Adjectives; CSLI	Adjectives list of sentential that-complements and extraposed that-complements	http://www.stanford.edu/group/csli_faust/Lexical_Resources/Polarity-Lexicon-of-Adjectives/	Copyright 2012 CSLI
Lexicon of Phrasal Implicatives; CSLI	Phrasal Implicatives work by Lauri Karttunen	http://www.stanford.edu/group/csli_faust/Lexical_Resources/Phrasal-Implicatives/	Copyright 2012 CSLI
LexBase; PARC	LexBase is a lexical database manager that reads the terms and the lexical and semantic relations defined by the WordNet system and stores them in a memory resident database that can be queried through the lookup of nouns, verbs, adjectives and adverbs as well as retrieving related words and concepts such as synonyms, antonyms, hypernyms, meronyms, etc. It also supports the programmatic editing of the database.	Developed by PARC; delivered to SRI. Contact Lynn Voss <loren.voss@sri.com>.	Modified BSD License

BD-1; PARC	BD-1 is an indexing system that provides a scalable, flexible database for retrieving complex linguistic structures. BD-1 is a database system for storing and querying of n-tuples. The system is designed specifically to provide efficient search & natural representations of annotated text. BD-1 can function as a key-value database, a triple store, or an n-tuple store. BD-1 is compatible with the Berkeley database, and also supports a query language for n-tuples that is a simplified subset of the SPARQL query language for RDF. It can be configured to use memory as a cache for its data store — which is particularly useful for lexical resources such as WordNet that can easily be accommodated in current machines.	Developed by PARC; delivered to SRI. Contact Lynn Voss <loren.voss@sri.com>.	Modified BSD License

APPENDIX C. ONYX FINAL REPORT

Overview

Ellipsis is a linguistic process that renders certain aspects of text meaning invisible at surface structure, thereby making them inaccessible to most current text processing methods. Ellipsis is considered one of the more difficult aspects of text processing and, accordingly, has not been widely pursued in NLP applications.[17] However, not all cases of ellipsis are created equal: some can be detected and resolved with high confidence within the current state of the art. We have been working toward configuring a system that can resolve one class of elliptical phenomena: elided scopes of modality.

We have addressed the problem of elided scopes of modality from two perspectives.

1. We developed a full microtheory of modal-scope ellipsis treatment that is being incorporated into the language-enabled intelligent agents in the OntoAgent cognitive architecture. This direction of work is reported in conference paper "Resolving Elided Scopes of Modality in OntoAgent" (McShane et al. 2012), which was presented at the First Annual Conference on Advances in Cognitive Systems (Dec., 2012) and is being delivered as part of this project work. This approach employs all of the static knowledge resources and reasoning engines available to OntoAgent intelligent agents.
2. We developed a method of detecting and resolving a subset of cases of modal scope ellipsis that can be applied to big data. In order to work over big data in real time, the approach uses only a subset of the resources and reasoners available in our environment and replaces some of the more resource-intensive aspects of processing with cheaper proxies. The goal was to focus on achieving high precision over a large data set.

Since the content and results of the first direction of work are well described in the cited paper, this report concentrates on the second direction of work.

[17] As Spenader & Hendriks (2005) write in the introduction to the proceedings of a workshop devoted to ellipsis in NLP, "The area of ellipsis resolution and generation has long been neglected in work on natural language processing, and there are few examples of working systems or computational algorithms." In fact, of the ten contributions to that workshop, only one reports an implemented system, the others discussing corpus studies of ellipsis, descriptive analyses of phenomena, or theoretical (typically, pragmatic) frameworks in which ellipsis might be treated.

Background

Even though big data sources such as the Worldwide Web and the Gigaword corpus (Graf and Cieri 2003)[18] contain a vast number of text strings, a significant portion of the *information* they contain is not represented in the surface text. One source of unexpressed information is the grammatical process of ellipsis, which is the non-expression of meanings that can be understood from the context. Whereas ellipsis increases the efficiency of natural language use by people, it introduces problems for machine processing. For example, the final clause in sentence (1) will be of little use to most knowledge extraction engines until and unless it is decorated with metadata indicating that what the toddler *could* do is *teach illiterate women to read and write.* (In this and subsequent examples, elided categories are indicated by [e] and their sponsors – i.e., the material used to recover their meaning – are surrounded by square brackets.[19])

(1) Aid workers in war-ravaged Kabul were stunned when a toddler from a poor family offered to [teach illiterate women to read and write] – and then promptly proved he could [e].

Although resolving some cases of ellipsis requires sophisticated semantic and pragmatic reasoning, not all cases are so difficult. The method we have developed to work over big data automatically *detects* which cases of ellipsis can be resolved with high confidence and treats only those cases.

The utility of work on modality, and envisioned application areas, are similar to those of the burgeoning field of sentiment analysis: permitting information extraction engines to separate fact from opinion; allowing summarization engines to distinguish what might happen to what might have happened from what did happen; helping intelligence analysts to detect intentions and threats; and so on.

A given proposition can be scoped over by many different types of modal meanings, as shown by the italicized strings in (2):

(2) The US *has <has not, might have, cannot have, should not, might, wants to, does not want to, seems to have, could not have, is believed to have, failed to***,** etc.> sign(ed) the treaty.

Within the NLP framework in which we work, called Ontological Semantics (Nirenburg and Raskin 2004), ten types of modal meanings are distinguished: epistemic, belief, obligative, permissive, potential, evaluative, intentional, epiteuctic, effort, and volitive. Each can be represented in language by various words and phrases: e.g., **volitive**: *want to, not want to*; **permissive**: *may, may not*; **epiteuctic**: *succeed in, fail to*. Each modal meaning has a scope that represents the event the meaning applies to. The scope can be overtly specified or elided; if elided, it must have a *sponsor* that allows the reader/hearer to reconstruct the meaning.

Sponsors for elided referring expressions – like sponsors for any referring expression – are most accurately understood as semantic entities, not text strings. The reason is that the elided category

[18] Available at http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T05.

[19] The inflectional form of the sponsor is irrelevant since, in reality, an elided *meaning* is recovered based on the *meaning* of the sponsor.

and its sponsor do not always stand in a precise coreference relationship. For example, in (3) the sponsor for [e] is a *different* instance of the event of sending that affects *different* children.

(3) Better-off parents could [send their children abroad for English education] but poorer families could not [e].

Inexact coreference like this is referred to in the linguistic literature as "sloppy identity" – as contrasted with "strict identity," by which two categories precisely corefer (Fiengo and May 1994). Due to the potential of sloppy identity, pointing to text strings as sponsors for elided categories is, at best, inexact. However, the alternative – generating full semantic representations for all inputs – is infeasible for big data in the near term, so the work reported here, which is oriented namely toward big data, uses text strings as a proxy for semantically analyzed sponsors.

Another drawback of pointing to text strings as sponsors is that it is not the case that the largest VP in the sponsor text chunk always resolves the ellipsis. For example, in (4) the modal element 'wanted to' is *ex*cluded from the ellipsis resolution, whereas in (5) it is *in*cluded.

(4) I at least wanted to [go three sets] if I could [e].
(5) I [wanted to go three sets] but he didn't [e].

In short, determining the general text span in which the ellipsis sponsor is located is not full-fledged reference resolution: other semantic decisions must be made as well. In the work reported here, we do not pursue the issue of strict vs. sloppy identity between elided categories and their sponsors, but we *do* pursue whether or not modalities in the sponsor text span should be included in, or excluded from, the actual sponsor.

Approach & Select Aspects of Evaluation

We describe our treatment of modal scope ellipsis by following the flow chart in Figure 10 (page 92), first by giving a top-level overview then by detailing the functioning of each engine.

The system takes as input our indexed version of the Gigaword corpus and selects examples that include modal scope ellipsis. Those examples are analyzed by a preprocessor and parser which, for purposes of this experiment, are treated as black boxes.[20] The next series of engines, which use heuristic evidence from preprocessing and parsing, act as sieves (cf., e.g., [130] for the sieve metaphor), each one catching examples of a particular profile to treat. The output of the sieves is a pointer to the text span that is believed to contain the sponsor. Once the system knows where to look for the sponsor, it needs to evaluate whether any modalities contained therein should be included in, or excluded from, the sponsor. This work is carried out by the Modality Evaluator. The output of this engine is a set of examples decorated with metadata indicating how to resolve the elided scope of modality.

We now consider the process in more detail, engine by engine.

[20] We use the Stanford preprocessor, supplemented by our own preprocessor, as well as the Stanford dependency parser (de Marneffe, MacCartney and Manning 2006).

Modal scope ellipsis detector. This engine extracts from the Gigaword corpus elliptical contexts of interest using the single pattern "*verbal modal element*" + "*period / semi-colon / comma*", which has low recall but quite high precision. Few false positives were detected, the most common ones being (1) the word "might" used as a noun at the end of the sentence: *"...marking the symbolic fall of the dictatorship to US might."*; and (2) the verb "did" being used as a full-fledged verb rather than an auxiliary: "*proud of what he did.*"

One can readily think of ways of improving recall, as by allowing an adverb to intervene between the modal element and the punctuation mark: *John didn't want to [e] either*. However, it is noteworthy that many patterns that we assumed would have high precision – such as "*verbal modal element*" + "*comma*" – returned such extensive false positives (*John didn't want to, for example, finish the painting*) that we chose to exclude them from our initial experimentation.

One of the main reasons we decided it was premature to publish this "big data" aspect of our work was that a formal evaluation would not be representative without a more robust detection method. That is, we used very constrained heuristics as an initial approximation to get the system up and running but did not have time in the project to return to the issue of detection and achieve better coverage. We hypothesize that coverage could be significantly improved, without extensive false positives, given a reasonably small additional development effort.

Sieve 1: Elliptical multi-word expression detector. A well-known method for improving text analysis overall is exploiting lexically recorded multi-word expressions in lieu of compositional analysis. Many configurations containing modal scope ellipsis can conveniently be lexically recorded as multi-word expressions that contain combinations of fixed and variable elements. An example is the adverbial *as [adv] as X can**, which covers contexts such as:

(6) Liz Mikropoulos of Bellaire, Ohio [climbed] as far as she could [e].

For this experiment we used the following multi-word patterns, but corpus evidence shows that the list could profitably be extended.

Table 3: Inventory of multi-word expressions containing modal scope ellipsis.

whatever/what [NP] *can	all the [NP] (that) [NP] *can
(anything and) everything (that) [NP] *can	as far as [NP] *MOD-WORD
wherever/where [NP] *can	as much/many as [NP] *MOD-WORD
in any place/anywhere [NP] *can	as [adj] as [NP] *MOD-WORD
in any way (that)/however [NP] *can	as [adv] as [NP] *MOD-WORD
as best (as) [NP] *can	as best (as) [NP] *MOD-WORD
whenever /when [NP] *MOD-WORD	as much/many [NP] as [NP] *MOD-WORD

Key: ( ) optional element; / an option; * any inflectional form; [NP] nominal [adj] adjective [adv] adverb; MOD-WORD: verbal indicator of modality.

The multi-word lexicon entries for each of these patterns indicates the syntactic structure that serves as the sponsor for the ellipsis: e.g., in (6), [e] is resolved by the verbal head that is modified by the multi-word adverbial.

In our evaluations of this engine, we found no false positives but are aware that false negatives can occur in cases of complex inputs: e.g., if a variable element was listed in our pattern as [NP] but the parser did not include a relative clause within the maximal projection of the [NP], then the engine would not recognize the pattern.

Sieve 2: Simple parallel configuration detector. This engine detects what we call “simple parallel modal scope ellipsis configurations”, as illustrated by (7)-(10).

(7) He encouraged his children [to take interest in the family business], and they did [e].

(8) Seven golfers, including Leonard, needed to [win] and didn't [e].

(9) They [managed to get out]; his wife did not [e].

(10) I at least wanted to [go three sets] if I could [e].

We will define what we mean by “parallel” and “simple” in turn.

“Parallel”. Each applicable context contains an ellipsis clause directly preceded by a conjunct that is syntactically connected to it in one of several highly constrained ways that can be loosely described as showing syntactic parallelism. The conjunct relationships that seem to have the strongest predictive power for modal scope ellipsis resolution are clausal coordination, verb phrase (VP) coordination, parataxis (juxtaposition using certain punctuation marks) and variations on the *if... then* theme (e.g., *if...* [no overt then]...; *if...when*; ... *if...*), as illustrated in turn by the examples (7)-(10) above.

The reason for exploiting syntactic parallelism to predict ellipsis resolution derives from the well-documented linguistic effects of parallelism (e.g., Goodall, 2009). The use of ellipsis tends to impose a greater cognitive burden on the interlocutor than an overt category would, and in order to fulfill the corresponding discourse obligation, the speaker can foster resolution by employing a highly predictive parallel structure.

“*Simple”.* The predictive power of parallel configurations decreases precipitously if the conjuncts – particularly the first – contain relative or subordinate clauses because such structures provide additional candidate sponsors for the elided verb phrase. For example, if we rewrite example (9) such that the first clause includes several embedded clauses, as in (11), it becomes necessary to carry out sophisticated reasoning about the world to determine which action the wife did not do: *arrive? cross the border? act quickly? manage to get out? all of the preceding events together*?

(11) They managed to get out because they acted quickly and crossed the border before the troops arrived; his wife did not [e].

Operationalizing “simple parallel”. We operationalized the notion of “simple parallel” configurations in terms of the output of the Stanford dependency parser. Applicable configurations contained exactly one instance of a CONJ, ADVCL or PARATAXIS dependency (which indicates “parallelism”, using our loose definition), and no instances of a CCOMP, PARTMOD, RCMOD, DEP or COMPLM dependency (which indicate various types of embedded structures that make a configuration not “simple”). For ease of reference, we refer to the first group as “whitelisted dependencies” and the second group as “blacklisted dependencies”. Table 4 shows examples of each type of dependency – not in an elliptical configuration) – with the blacklisted ones having a gray background.

Table 4: Whitelisted and blacklisted dependencies.

Examples/Description	Dependencies
Sue dove and caught the frisbee.	conj(dove, caught)
If you wash, I'll dry.	advcl(dry, wash)
You wash; I'll dry.	parataxis(wash, dry)
He said that she was swimming.	ccomp(said, swimming)
Bonds bought by investors came due.	partmod(bonds, bought)
I saw the house you bought.	rcmod(house, bought)
He said that she was swimming.	complm(swimming, that)
a catch-all for uncategorized dependencies	dep(x, y)

Although simple parallel configurations do not represent a large percentage of elliptical examples in the corpus, when they are found, they offer very confident predictions about which conjunct contains the ellipsis sponsor.

Sieve 3. Less-simple parallel configuration detector. In an attempt achieve greater recall (i.e., coverage of the corpus) while still leveraging the predictive power of structural parallelism, we experimented with various methods of relaxing the definition of "simple". The options were to permit examples to have more than one whitelisted dependency, or to have one or more blacklisted dependencies. Although we tried to determine which dependencies would have the least detrimental effects on predictive power, our experimentation yielded disappointing results. We attribute this loss of predictive power given any "additional" dependencies to at least three factors. First, as mentioned earlier, ellipsis judgments should ideally be made with the contribution of semantic analysis. Using syntax as a proxy for semantics only works well when the syntactic structure is so simple that no matter *what* the text means, the sponsor must be in a given structural correlation with the elided category. Second, the results of parsing were not always as expected; but, since this effort to expand big data only makes sense if the system works in fully independent mode, we did not engage in manually correcting parses. Finally, certain ideas about how to ignore irrelevant dependencies turned out to be difficult to operationalize, such as the desire to prune off the initial clauses of very long sentences on the hypothesis that the sponsor for a sentence-end ellipsis would be located toward the end of the sentence.

We included in the evaluation just two of the less-simple configurations with which we experimented. In the first, we permitted contexts to have exactly one blacklisted dependency of the type DEP, both of whose arguments were elements of the first conjunct. For example, in (12), the dependency that establishes the parallelism used to resolve the ellipsis is (ADVCL (could, fails)), and the "extra" blacklisted dependency that is ignored – and does not impede the system's correct ellipsis resolution – is (DEP (fails, if)).

(12) Even if political pressure fails to [deliver money to some of the 20,000 people forced to evacuate because of the fire], the federal courts could [e].

The second relaxation of "simple" that we formally evaluated permitted contexts to have exactly one extra instance of parataxis, one of whose arguments was in the hypothesized sponsor conjunct and none of whose arguments was in the ellipsis conjunct. For example, in (13), the dependency that establishes the parallelism used to resolve the ellipsis is (ADVCL (kept, could)), and the "extra" whitelisted dependency that is ignored – and does not impede the system's correct ellipsis resolution – is (PARATAXIS (forced, kept)).

(13) Crusader forced to confess: Spitzer [kept prostitution shame to self] until he no longer could [e].

We chose to evaluate these two particular relaxation strategies because, on initial examination, they seemed to work pretty well. It is entirely possible that additional corpus study, combined with an intimate understanding of the Stanford parser's dependency decisions, could yield additional useful relaxations to our baseline definition of "simple".

Sieve 4. Nearest clause & modality detector. Unlike sieves 2 and 3, this one does not seek "parallel" syntactic structures. Instead, it seeks contexts in which the most proximate preceding clause – no matter its relationship with the ellipsis clause – is also scoped over by modality. The rationale for this sieve is that modal clauses often serve as sponsors for elided scopes of modality, and recency is a strong vote in favor of candidate sponsors for all kinds of reference resolution. For example, in (14) the most proximate clause is *would like to find his own lawyer*, which includes the modal element *would like to*.

(14) Brandon said he would like [to find his own lawyer] but was not sure he could [e].

This sieve selects "would like to find his own lawyer" as the sponsor *conjunct* after which the Modality Evaluator, described below, will correctly exclude the modal element from the *actual sponsor*.

Sieve 5. Nearest modality detector. This engine walks back through the text strings – ignoring the syntactic parse – in search of the first verbal modal element it encounters. If it finds one, it selects that element's conjunct (determined using the syntactic parse) as the sponsor conjunct. This engine is a last ditch effort to use leverage *some* heuristic evidence (apart from just recency, which is the final default for ellipsis resolution) to select a sponsor conjunct. We are trying to exploit the generalization that if a text, e.g., contains "succeed [e]" there is a good chance that the sponsor will include "try to X", and it would be nice to find that instance of "try to X" if it exists. As will be discussed later, the key to getting useful results from this method lies in treating only select pairs of modalities this way.

At this point, our sieves have selected all of the examples that comply with their heuristic filters, and they have pointed to the conjunct in which they believe the sponsor is located. As discussed earlier, selecting the sponsor conjunct does not exhaust the work of selecting the actual sponsor. The next engine in the pipeline will carry out one aspect of that work: if the sponsor conjunct contains modal elements, it will determine whether they should be included in, or excluded from, the sponsor.

Modality Evaluator. Any proposition that serves as a sponsor for ellipsis resolution could, itself, be scoped over by modality, in which case the modality might need to be *in*cluded in or *ex*cluded from the ellipsis resolution. For example, the sponsor conjuncts in (15)-(17) all contain modal meanings scoping over the main proposition; however, whereas the modalities in first two are *ex*cluded from the ellipsis resolution, the modality in the last one is *in*cluded in the resolution, as detailed by the description following each example.

(15) The media also blasted Erjavec for taking his wife with him on a trip and insisted he should have [gone through the customs] as all citizens must [e].

[e] = ‘go through the customs’, *not* ‘should go through the customs’

(16) On Friday night, he wanted to [go out in style], and he did [e].

[e] = ‘go out in style’, *not* ‘want to go out in style’

(17) The scheduled train [managed to stop in response to frantic radio warnings], but the supplementary train didn't [e].

[e] = ‘manage to stop in response....’ *not* ‘stop in response...’

We have formulated a *preliminary* set of rules to predict whether modality should be included in or excluded from ellipsis reconstruction. The key feature in these rules is “modality-correlation-strength”, whose values are “match” (the modality *type* in both conjuncts matches), “strong” (there is a frequently encountered correlation of modalities, such as *tried to... couldn’t; tried to... succeeded; wanted to... couldn’t*), and “weak” (there is no special correlation between the modality types, as in *wanted to... didn’t have to*). Table 5 shows some of our “strong” correlation rules, along with toy examples selected for clarity of comparison.

Table 5: Sample Modality Correlation Heuristics

The sponsor clause contains	**The ellipsis-licensing modality is**	**Include in the ellipsis resolution**	**Ex**
Effort or *volitive*	*Epiteuctic* or *epistemic*	Only the scope of the modality	18
Obligative or *volitive*	*Potential*	Only the scope of the modality	19
Any types of modality; the outer one is not epistemic (negation)	*Epistemic* only (e.g., *did, didn’t*)	All modalities along with the scope	20
Any number of modalities; the outer one is epistemic (negation)	Epistemic only (*do, don’t, etc.*)	All modalities except the outer epistemic	21

(18) John wanted to ski and did [ski].

(19) John had to ski and could [ski].

(20) John wanted to try to ski but Mary didn’t [want to try to ski]

(21) John didn’t want to try to ski but Mary did [want to try to ski]

We discuss known limitations of these rules, and suggestions for improving them, in the evaluation below.

The output of the ellipsis processing system is a set of elliptical examples and their sponsors. This information could readily, automatically, be converted into metadata supplementation of any data, including big data.

Preliminary Evaluation

We carried out several rounds of evaluations that proved useful to guide iterative system development.

Inputs analyzed as “simple parallel configurations” were divided into 16 classes, each input classified by (a) one of the four types parallelism (ADVCL, CONJ-CL, CONJ-VP, PARATAXIS) and (b) one of the four outcomes of modality correlations across conjuncts (no modality in the first conjunct; modality matching; strong modal correlation; weak modal correlation). For each example, the outcome of automatic ellipsis resolution could be: resolution was correct, which included selecting the correct verbal head as the sponsor and deciding whether 1st-conjunct modalities should be included in the ellipsis resolution; the correct head was selected but modal treatment was incorrect; the wrong head was selected; “other” (e.g., the context was uninterpretable to the human evaluator). Inputs analyzed as “parallel configurations permitting one DEP dependency” and “parallel configurations that permitted one “extra” PARA dependency” were evaluated using the same metrics. Inputs treated by sieves 4 and 5 were divided into three groups each, representing the three possible modal correlations. The same 4 outcomes for the evaluation of each example were employed for these sieves. Some useful generalizations resulted from these preliminary evaluations.

The “simple parallel configurations” yielded very high precision in detecting the sponsor conjunct, but the modality correlation rules sometimes failed to correctly determine whether to included or exclude the sponsor-clause modality in the actual sponsor.

One type of heuristic evidence that we believe could not only substantially improve the work of the modality evaluator but also prove useful in other ways is reference resolution for the subject of the ellipsis clause. For example, if our ellipsis engines had access to metadata indicating that “they” in (22) referred to “tragedy, betrayal and human suffering”, then the engines could exploit a new type of parallelism: parallelism between the subject of the sponsor clause and the subject of the ellipsis clause. More analysis would be needed to determine how much predictive power subject matching could afford, and how much an incorrect subject resolution would detract from ellipsis resolution.[21]

(22) In his foreword, Stephen R. Treat, director of the Penn Council for Relationships in Philadelphia, calls Mazo's book “an important contribution because tragedy, betrayal and human suffering [exist], no matter how much we wish they [=”tragedy, betrayal and human suffering”] didn't [e].”

As mentioned earlier, reference resolution of the ellipsis-clause subject could also be incorporated into an improved inventory of modality-correlation rules, examples of which are presented in Table 4. That is, predictions about whether a modal element should be included in or excluded from the ellipsis resolution are affected by whether the subject is the same or different in the ellipsis- and sponsor-conjuncts. We did not anticipate that our initial set of modal-correlation rules, which is quite coarse-grained, would perform exceptionally well – and,

[21] Our early experiments did include reference resolution of ellipsis-clause subjects, but we decided to exclude that process for the time being in order not to get distracted by work on improving the accuracy of our more comprehensive reference resolution engine.

indeed, it did not. However, we believe that rules of this type have the potential to be quite useful given additional corpus analysis and certain targeted additional sources of heuristic evidence.

A frequent source of errors involved syntactic structures that included XCOMP and INFMOD dependencies. These dependencies, like many of the dependencies that are "blacklisted" for purposes of "simple parallel configurations", indicate the presence of main and subordinate verbal structures, either of which could head the ellipsis sponsor. For example, (23) contains an INFMOD dependency between 'sell' and 'came', either of which could – formally speaking – head the ellipsis sponsor: i.e., *he did come* or *he did sell it.*

(23) He came to Toronto to sell it and he did [e].

The reason we could not add XCOMP and INFMOD dependencies to our blacklist was that these dependencies are also used to indicate the relationship between modal elements and their scopes. What we need to do to improve the algorithm, therefore, is look not only at dependency *types* but also at the nature of their *arguments*: i.e., XCOMP and INFMOD dependencies that take a modal as one of their arguments can be treated as "simple parallel configurations", but ones whose arguments are both non-modal cannot. We can also attempt to treat some specific kinds of the non-modal instances of XCOMP and INFMOD. For example, verbs indicating requests tend to be *excluded* from ellipsis sponsors, and the predictive power is increased if the theme of the request is understood as coreferential with the subject of the ellipsis clause, as in (24).

(24) But honestly, if you ask me to [name two groups], I couldn't [e].

Finally, the precision of examples caught by Sieves 4 and 5 was not very high, but we have an ideas about how to improve it. What we were trying to capture by these sieves were highly predictive modality correlations like *tried to X ... failed [e]; wanted to X ... couldn't [e].* We need to reconfigure the Modality Evaluator, which runs after the sieves, to treat only those examples that show highly predictive modality correlations, relegating all others to the "untreated" batch.

The final observation about this evaluation is that an important aspect of pursuing a "lightweight"[22] approach to a more fundamental problem is knowing when to stop. The treatment of ellipsis, like any reference phenomenon, benefits from semantic and pragmatic analysis. Syntactic proxies for semantics are useful only in some cases, and spending undue effort trying to turn semantics into syntax for the more difficult cases is neither theoretically nor practically justified.

[22] This approach is considered lightweight because a preprocessor and parser could be used out of the box, and because a "heavyweight" approach would involve semantic analysis. The approach would not be considered lightweight for a language for which these such engines were not available or did not produce results of reasonable accuracy.

References for Appendix C

Fiengo, R., May, R.: *Indices and Identity*. Cambridge, Mass.: The MIT Press (1994).

Graff, D., Cieri, C.: English Gigaword. Linguistic Data Consortium. Philadelphia (2003)

de Marneffe, M, MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. *Proceedings of LREC* (2006)

Goodall, G.: *Parallel Structures in Syntax: Coordination, causatives and restructuring*. Cambridge University Press (2009)

McShane, M., Nirenburg, S., Beale, S. and Johnson, B.: Resolving elided scopes of modality in OntoAgent. *Advances in Cognitive Systems* (2012)

Nirenburg, S., Raskin, V.: *Ontological Semantics*. The MIT Press (2004)

Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2*, 1–135 (2008)

Spenader, J., Hendriks, P. Combining multiple information sources for ellipsis, Introduction to Special Issue in *Research on Language and Computation*, *4,* 327-333 (2006)

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

APPENDIX D. IHMC FINAL REPORT

Processing Modality and Related Phenomena in Machine Reading

By Institute for Human and Machine Cognition (IHMC, Prof. Wilks)

Summary

IHMC executed an exploratory project to locate proto-beliefs of individual *Ummah* message board posters on a large scale. These beliefs could then be examined to determine the consistency of an individual poster's beliefs and to identify where that individual's beliefs conflict with the beliefs of others; such conflicts of belief could occur either within the context of a single thread or in the context of all threads.

Information Flow

In the information flow of the completed system, facts were extracted from the *Ummah* message board postings using unsupervised methods for information extraction. These facts were then linked to individual posters as beliefs or assertions in a belief management engine. Finally, heuristics were used to investigate confirmations and negations of beliefs within and outside individual message threads.

As an exploratory effort, the project's aim was to determine the feasibility of our approach to extraction and comprehension of agents' interrelated beliefs. The primary outcome of the completed work is a positive demonstration of the extraction of these beliefs. In particular, we demonstrated that 1) beliefs could be extracted from the unstructured data contained in an online forum, 2) represented in the ViewGen belief engine, and 3) scored using heuristic approaches similar to the FactRank (Jain & Pantel, 2010) algorithm.

With the successful conclusion of this exploratory project, the research effort is being extended and expanded as part the DARPA DEFT project. In this report, we provide some motivating background and then detail the technical tasks and activities comprising the conducted research.

Background

We suggested some years ago in the context of work on belief systems (Ballim & Wilks, 1991*a*) that the relative maturity of Information Extraction (IE) systems (Ciravegna & Wilks, 2003) now provided a route for the large-scale population of belief computation systems, and a way out of the toy scale of such work and into large-scale, evaluable, NLP. Since then, large-scale unsupervised fact harvesting from text has been pioneered by Pantel as an alternative to more focused, classic Information Extraction. (Both of these threads can be seen as part of the core of *machine reading*.) Pantel has also proposed simple heuristics (Jain & Pantel, 2010) to check the consistency of assertions and so prune out false positive "facts," but this was done in areas (such as: who were the directors and actors in such-and-such films) where the predicates involved are straightforward in a way that discussions in, say, a Muslim-orientated English blog like *Ummah*, would not be. With these observations as impetus, we proposed to apply exploratory techniques for assessing the interaction of beliefs in online message board threads where the thread span gives some constraint on the spread of a discussion for this purpose.

Tasks and Activities

Our approach was to decompose the overall research into three parallel tasks:

1. Fact extraction from the *Ummah* message board corpus
2. Belief representation and point-of-view ascription via the ViewGen belief engine
3. Heuristics for belief consistency and confirmation

The work performed on these three tasks is summarized in the following sections.

Fact extraction from the Ummah message board corpus

IHMC extracted the proto-beliefs of posters participating in online discussion forums. Working upwards from the corpus a public Muslim message board (*Ummah*) of about 1.5 million words — we populated a database organized by the originating individual poster, such that the beliefs/facts expressed in each post when extracted can be assigned to that posting. To implement this pipeline, we acquired the *GATE* (General Architecture for Text Engineering) platform that Wilks originally developed at Sheffield and a number of lexical resources to begin work on large-scale proposition extraction.

We built a GATE pipeline that includes the usual steps: tokenization, sentence splitting, POS tagging and noun/verb phrase chunking. We also incorporated Webb et al.'s (2005) DAT tagger in GATE as a plugin so the Dialogue Act information is available in GATE annotations.

We investigated a number of off-the-shelf semantic parsers and found two packages that claim to produce predicate/argument information of a sentence. The first was Predicate-Argument eXtractor (PAX) by Krestel, et al. (2010) and the second one is multilingual semantic role labeling by Björkelund, et al. (2009). We found both packages produced good results on simple and short sentences but fail on long and complex ones. In addition, we implemented the triplet extraction algorithm by Rusu, et al. (2007) using output of Stanford parser. We found this approach gave better results and more flexible as it can be easily extended. To tackle problems caused by complex sentences, we developed a simple 'chunker' plugin for GATE to chop long sentences into parts based on heuristics derived from a set of manually annotated examples. This gave further improvement of our triplet extraction algorithm.

A comparison follows of the output of the different analysis systems on three of the gold standard *Ummah* posts annotated by us. For each sentence, there are four outputs: Gold, PAX, SRL, and SP. PAX is the triple extraction GATE plugin, SRL is a semantic role labeler, and SP is a triple extractor based on a few rules. By analyzing the syntax tree of Stanford parser, the results are close to the Gold annotations on short sentences that have simple structure. None of these approaches worked on long and complicated sentences of the kind found in the blog.

1. "Dhakiyya, maybe you said something and didn't realize what you just said."

   Gold: ADDRESSEE ##NOT+REALISE ##WHAT+SAID

   PAX: s:you v:said [Negated] o:something
   s:you v:said [Negated] o:what

   SRL: s:you v:said o:something and didn't realize

SP: s:Dhakiyya v:said o:something
s:Dhakiyya v:didn't realize o:you said
s:you v:said o:

2. "I would think all jews are Zionists in one form or another."

   Gold: ALL+JEWS###BE###ZIONISTS

   PAX: s:jews v:Zionists o:one form another
   s:I v:think o:

   SRL: s:I v:think o:all jews are Zionists in one form or another

   SP: s:jews v:are o:Zionists
   s:I v:think o:jews are Zionists

3. "It is a means and methods about how to bring zionism about is the difference between a declared zionist jew and an non declared one."

   Gold: DECLARED+ZIONIST+JEW###DIFFERENT+FROM
   ###NONDECLARED+ZIONIST+JEW

   PAX: s:zionism v:difference o:a declared zionist jew an non one
   s:It v:means o:zionism about the difference a declared zionist jew an non one
   s:non v:declared o:one

   SRL: s: v:bring o:zionism
   s:an non v:declared o:one

   SP: s:means methods v:bring o:
   s:It v:is o:means methods
   s:non v:declared o:
   s:zionism v:is o:difference

To tackle with problems caused by complex sentences, we developed a simple 'chunker' plugin in GATE to chop long sentences into parts based on heuristics derived from a set of manually annotated examples. This gave further improvement of our triplet extraction algorithm. After the initial syntactic extracted forms given above, the SP representations of sentences are converted to triple form with predicates, as in:

maybe you said something and didn't realize L2
ADDRESSEE – SAID – SOMETHING
ADDRESSEE – NOT+REALIZE – L2

what you just said.
ADDRESSEE – SAID – L2

An initial hand-coded set of 20 sentences was produced in this form and the algorithm was run on an unseen set of 10 test sentences, with a resulting precision of 80% for the test set.

Our future work will include the development of more sophisticated sentence splitters and reconstructors than those available in GATE or elsewhere, while trying to remove "junk" from the colloquial forms in the message board posts so that more core parts of the sentence remain for analysis.

Belief representation and point-of-view ascription via the ViewGen belief engine

The extracted beliefs of posters are represented using the ViewGen (Wilks, 2011; Ballim & Wilks 1991*b*; Ballim et al., 1991) paradigm. Briefly, the ViewGen paradigm is a particular theoretical approach to belief representation and ascription, specifically for dealing with multiple points of view (usually an agent's beliefs about itself and others, including the various agents' beliefs about objects and events) in the presence of complexities such as *de dicto*, *de re*, and *de se* distinctions, metaphors, and conflated entities. Note that while there have been several experimental implementations of the ViewGen paradigm there is no canonical implementation of the ViewGen belief engine.

In the course of this project, we implemented two versions of the ViewGen belief engine. The first version was written in *Prolog* with a *Redis* key-value store backend. This first version was the primary proving ground for our work. The second version is still underway and it will continue in development as part of the DARPA DEFT program; it is written in *Java* and *Clojure* with *MySQL* and *PostgreSQL* backends. The differences between these two versions are a matter of performance and scalability rather than functionality. (The second implementation — Java/Clojure with relational database backend — should scale to very large corpora and integrate well with very large scale DBMS; however, testing scalability was beyond the scope of our exploratory work.) Thus, the following description of ViewGen functionality can be applied to either implementation.

In operation, ViewGen partitions the space of beliefs according to agents and topics. This organization can be viewed conceptually either as a database indexing agent and topic to beliefs held by the agent about the topic, or as a tree of *agent viewpoints*[23] and *topic environments*[24] rooted in the system's own viewpoint (see the figure below). With respect to the *Ummah* message board, each poster has a corresponding viewpoint under the system's viewpoint. These viewpoints are populated with topic environments and extracted beliefs — those being the output of the fact extraction process described earlier.

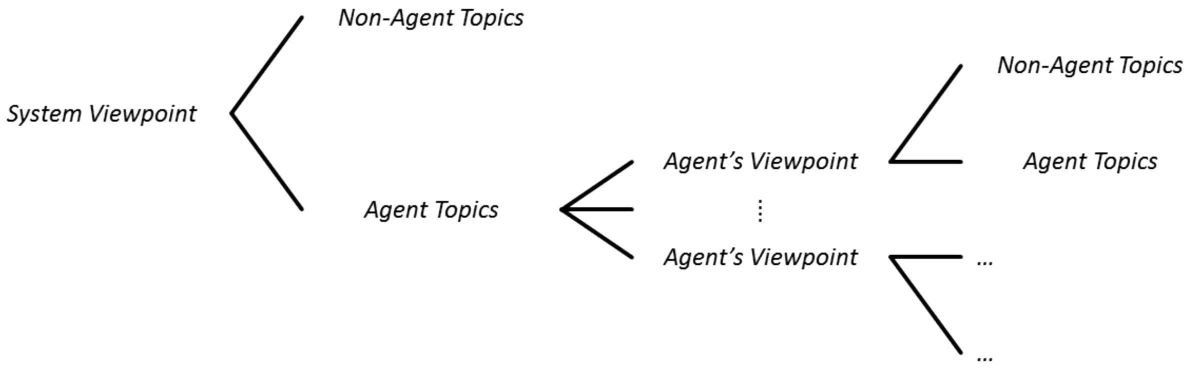


Figure 11: Topology of Agent Viewpoints and Topic Environments

[23] An *agent viewpoint* represents a specific agent's perspective, and it scopes lower viewpoints and topic environments as belonging to the agent (i.e., as being from the specific agent's point of view).

[24] A *topic environment* is a collection of topically related beliefs.

ViewGen uses an ascription algorithm for default reasoning to ascribe the system's (or an agent's) beliefs to other agents unless there is evidence to prevent it; evidence such as: (i) a contradiction between system beliefs and what the system believes are another agent's beliefs, and (ii) atypicality of belief including self-knowledge, belief competency, and expertise. That is to say, the algorithm proceeds on the assumption that another agent's beliefs are equivalent to the system's own beliefs *except where there is explicit evidence to the contrary*. Because of this assumption, the system only needs to store its own beliefs and any beliefs the system presumes others hold that conflict with the system's own or would otherwise not be ascribed (e.g., private self-knowledge).[25] Topic environments organize an agent's beliefs by relevance; they are used to limit the scope of reasoning and ascription to only those beliefs that are relevant to a given query or task. Thus, the system can efficiently determine, say, what agent *A* believes agent *B* believes about topic *T* without having to consider the totality of all the things that *A* believes and all the things that *A* believes that *B* believes.

The ViewGen ascription algorithm proceeds by generating a fixed sequence of topic environments that represent all of the relevant topics necessary to flesh out (via ascription) the total contents of a particular topic environment (e.g., the system's beliefs about agent *A*'s beliefs about agent *B*'s beliefs about topic *T*). For each environment in this sequence, the beliefs in that environment (source environment) are ascribed to the next environment in the sequence (target environment) unless the ascription of a particular belief is blocked, with this process resulting in a final environment populated with both explicit and ascribed beliefs.

With respect to FAUST, we investigated several algorithms for determining whether the ascription of a particular belief ought to be blocked. For all of these algorithms, the criterion for blocking an ascription was that the ascription would produce a new contradiction of beliefs in the resultant environment — this codifies the ViewGen principle that while an agent can hold contradictory beliefs, the ascription process should not ascribe new contradictions of belief to an agent. In addition, all of the algorithms made use of the *LeanCoP* theorem prover (Otten, 2010), which was integrated into the ViewGen belief engine.

The first algorithm checked that an individual belief in the source environment was logically consistent with each belief in the target environment (independent of other beliefs in either environment). This algorithm was fast and scaled well; however, it critically depends on the assumption that each belief in an environment is logically independent of other beliefs in the environment. While this assumption is a principle of the ViewGen paradigm, it is difficult to achieve in the context of automated belief extraction. With this difficulty in mind, the second algorithm checked that an individual belief in the source environment was jointly consistent with all of the beliefs in the target environment. Variants of this third algorithm also dealt with the issue of preexisting belief contradictions in the target environment. These variants decomposed an inconsistent target environment into consistent belief subsets with self-contradictory beliefs (i.e., inconsistent singleton belief sets) being ignored for the purpose of ascription. The variants differed in whether they were greedy or exhaustive algorithms. A fourth algorithm (and variants thereof) treated both the source and target environments as sets of logically related beliefs and attempted to ascribe the "best" consistent subset of beliefs from the source environment to the

[25] Note that agents are also valid topics, thus ViewGen has no difficulty with the fact that, say, the system believes that John is unintelligent and at the same time, the system believes that John believes that he himself is a genius.

target environments. Variants of this fourth algorithm decomposed both source and target environments into consistent belief subsets in similar fashion to variants of the third algorithm. Finally, a fifth algorithm (based on the greedy variant of the fourth) was developed which used a "confidence" metric to guide the greedy selection of belief subsets.

As one might expect, the described algorithms suffered from declining throughput performance due to the inherent computational complexity of consistency checking and subset generation. This performance degradation was only partly ameliorated by metric-driven, greedy variants. A conclusion drawn from this work is that there ought to be an additional processing step (prior to ascription and query in ViewGen) where the beliefs in each environment are reorganized; specifically, logically related yet independently represented beliefs ought to be combined into and replaced with a single belief/assertion. Conversely, any independent clauses of a belief ought to be detached and replaced with separate independently represented beliefs. This new processing step will enable the system to use the simplest and most efficient ascription algorithm (i.e., the first algorithm described above) by 'reorganizing' each environment into a set of weakly related, independently represented beliefs.

Heuristics for belief consistency and confirmation

Pantel's (Jain & Pantel, 2010) FactRank random walk algorithm was adapted for scoring belief consistency, in part because of its ability to handle noisy data in large, uncurated fact collections. Extracted beliefs can be structured in *n*-ary typed relations in a similar fashion to that used by Pantel, such as *acted-in<movie, actor>*. As with FactRank, a belief is strengthened when multiple relations assert the belief, and incorrect ascribed beliefs will appear less frequently in the extracted text than those legitimately held.

FactRank is built on a graph data structure where nodes are parameterized fact instances, and edges link instances that share parameters. Once a graph is constructed, other ranking algorithms such as PageRank (Page et al., 1999) can be applied to attempt to confirm an ascribed belief in the presence of inconsistent data.

Visual Depiction of a (Small) Graph of Beliefs as Scored by FactRank

We have developed several experimental algorithms for integrating the FactRank fact confirmation algorithm into ViewGen's core ascription algorithm. Our initial 'best subset' algorithm, which ascribes the highest scoring subset of consistent beliefs, performs well with beliefs are sparse but is not tractable when the belief space is dense. We have integrated FactRank as a scoring metric in our 'greedy' ascription algorithms, which were described in the previous section. While we have learned a lot about the use of random-walk scoring algorithms, several questions remain that are of importance to our use of such algorithms in a 'belief confirmation' context:

1. How do we properly score contradictory facts (say, *P* and *not-P*) versus the simple falsity (or non-confirmation) of a fact (say, that *P* is not true or not confirmed)?
2. In the context of beliefs and differing viewpoints, can beliefs be scored en masse regardless of viewpoint or should the beliefs of one agent be scored independently and in isolation from the beliefs of other agents?
3. Given that the score individual facts/beliefs is based on the overall graph, how brittle are rankings to topological changes — specifically, how do the ordered rankings of facts within an arbitrary sub-graph compare to a rescoring of that sub-graph as its own independent graph?

Concluding Remarks

During Phase 3 of the FAUST project, IHMC prototyped a system capable of extracting, modeling, and scoring beliefs assigned to forum posters and for representing posters' reflexive beliefs of themselves and others. This work will be continued in the DARPA Deep Exploration and Filtering of Text (DEFT) program where we will target a deep and robust analysis of multi-party conversations. We will continue maturing belief extraction and representation technologies that help expose pragmatic knowledge in a conversation (knowledge that is otherwise contextually bound and only implicitly expressed). We also plan to augment our existing capabilities with algorithms for explicating conversation dynamics (e.g., topic sequencing and sociolinguistic features) which will help inform the extraction, recognition, and projection of interlocutors' beliefs and intentions, and changes thereto over time.

References

Ballim, A., and Wilks, Y. (1991*a*) Beliefs, Stereotypes and Dynamic Agent Modeling. *User Modeling and User-Adapted Interaction* **1**(1): 33–65.

Ballim, A., and Wilks, Y. (1991*b*). *Artificial Believers: The Ascription of Belief*. Lawrence Erlbaum Associates.

Ballim, A., Wilks, Y., and Barnden, J.A. (1991). Belief Ascription, Metaphor, and Intensional Identification. *Cognitive Science* **15**(1): 133–171.

Björkelund, A., Hafdell, L., and Nugues, P. (2009). Multilingual semantic role labeling. In *Proc. of the 13th Conf. on Computational Natural Language Learning: Shared Task*, pp. 43–48.

Ciravegna, F., and Wilks, Y. (2003). Designing Adaptive Information Extraction for the Semantic Web in Amilcare, In *Annotation for the Semantic Web*, pp. 112–127. IOS Press.

Jain, A. and Pantel, P. (2010). FactRank: Random Walks on a Web of Facts. In *Proc. of 2010 Conf. on Computational Linguistics*, pp. 501–509.

Krestel, R., Witte, R., and Bergler, S. (2010). Predicate-Argument EXtractor (PAX). In *Proc. of the 2010 Conf. on Language Resources and Evaluation, Workshop on New Challenges for NLP Frameworks*, pp. 51–54.

Otten, J. (2010). Restricting Backtracking in Connection Calculi. *AI Communications* **23**(2–3): 159–182.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the Web. *Technical Report 1999/66*. Stanford University, Computer Science Department.

Rusu, D., Dali, L., Fortuna, B., Grobelnik, M., and Mladenic, D. (2007). Triplet extraction from sentences. In *Proc. of the 2007 Conf. on Data Mining and Data Warehouses*.

Webb, N., Hepple, N., and Wilks, Y. (2005). Dialogue Act Classification Based on Intra-Utterance Features. *In Proc. of the AAAI Workshop on Spoken Language Understanding*.

Wilks, Y. (2011). Protocols for Reference Sharing in a Belief Ascription Model of Communication. In *Advances in Cognitive Systems: Papers from the 2011 AAAI Fall Symposium*, pp. 337–344. AAAI Press.